In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.api as sm
from sklearn.model_selection import train_test_split as tts
from sklearn.metrics import r2_score
from sklearn.linear_model import LinearRegression as LR
from sklearn.metrics import mean_squared_error as MSE
from sklearn.model_selection import cross_val_score as CVS
from sklearn.preprocessing import PolynomialFeatures as PF
from sklearn.decomposition import PCA
from sklearn.feature_selection import RFE
from statsmodels.formula.api import ols
from statsmodels.tools.eval_measures import rmse as RMSE
import scipy.stats as stats
import numbers
import pickle
import folium
```

Modelling

Modelling requires several different operations. For ease of repetition, I decided to make functions for these operations, with different options for each which could even be used together.
The different operations and their functions are:

- Transform the Data
  - polynom
  - log
  - minmax_plus
  - norm
  - ohe
  - bin_latlong
  - -bin_basement
- Feature Selection
  - simple_selector
  - stepwise_selector
  - rfe_selector
- Handling Multicollinearity
  - multicoll_remove
- Create the Model
  - model
- Measuring Results
  - metrics
  - unlog

polynom

This function creates polynomial columns for the features passed in as to_poly. If poly is 'all' it also creates columns for multiplications between the columns passed in. Lastly, though it gets column names from poly.get_feature_names, these have to be altered to make it readable by statsmodels' ols function.

In [2]:
```
def polynom(data_t, to_poly, poly='all'):
    poly_order = 2
    if poly == 'all':
        poly=PF(poly_order)
        data_poly = poly.fit_transform(data_t[to_poly])
        data_poly = pd.DataFrame(data_poly)
        data_poly.columns = poly.get_feature_names(data_t[to_poly].columns)
        data_t = pd.concat([data_t.drop(to_poly, axis=1),
                            data_poly.drop('1',axis=1)], axis=1)
        data_t.rename(lambda col: col.replace('.0', '').replace(' ','__')
                      .replace('^','_pow_'), axis=1, inplace=True)
    elif poly == 'singles':
        for feat in to_poly:
            df = pd.DataFrame(data_t[feat])
            print(feat)
            poly=PF(poly_order)
            data_poly = poly.fit_transform(df)
            data_poly = pd.DataFrame(data_poly)
            data_poly.columns = ['drop', feat, feat + '_squared']
            data_t = pd.concat([data_t.drop(feat, axis=1),
                                data_poly.drop('drop',axis=1)], axis=1)
    return data_t
```

log

log simply replaces the data with the log of the data for the columns passed in as to_log

In [3]:
```
def log(data_t, to_log):
    for feat in to_log:
        data_t[feat] = data_t[feat].map(lambda x: np.log(x) if x!=0 else 0)
        #Note 'if' included so that 0 values wouldn't error, will still error
        #...between 0 and 1. Also, implies that the original 0 value was a 1
        #...but as these are around values of 100s, 1000s, the effect is minim
al
    return data_t
```

norm

For each column in to_norm, if this isn't the test data, norm calculates its mean and stdev, using them to normalise the column's values. It then records the statistics used for each column so they can be used later for the test data.
If this is test data (test=True) it reads the stats from the record made previously

In [4]:
```
def norm(data_t, to_norm, test=False):
    stats = {}
    if test:
        with open('norm_stats.pickle', 'rb') as f:
            stats = pickle.load(f)
    for feat in to_norm:
        ft = data_t[feat]
        if test:
            mean = stats[feat][0]
            stdev = stats[feat][1]
        else:
            mean = np.mean(ft)
            stdev = np.std(ft)
        data_t[feat] = (ft-mean) / stdev
        #record stats used:
        stats[feat] = [mean, stdev]
    if not test:
        with open('norm_stats.pickle', 'wb') as f:
            pickle.dump(stats, f)
    return data_t
```

minmax_plus

This function performs min-max scaling on the columns in to_minmax, and also multiplies by 100, and adds 1 to,each value to make them log-able. It records the stats used for the train-data and reads it for test-data, in the same way as norm.

In [5]:
```
def minmax_plus(data_t, to_minmax, test=False):
    stats = {}
    if test:
        with open('minmax_stats.pickle', 'rb') as f:
            stats = pickle.load(f)
    for feat in to_minmax:
        ft = data_t[feat]
        if test:
            min_ = stats[feat][0]
            max_ = stats[feat][1]
        else:
            min_ = ft.min()
            max_ = ft.max()
        data_t[feat] = 1 + 100*(ft-min_)/(max_-min_)
        #record stats used:
        stats[feat] = [min_, max_]
    with open('minmax_stats.pickle', 'wb') as f:
        pickle.dump(stats, f)
    return data_t
```

ohe

ohe gets dummies for the columns in to_ohe, drops the original, and adds these columns to the dataframe. Lastly, the column names have to be altered to make it readable by statsmodels' ols function.

In [6]:
```
def ohe(data_t, to_ohe):
    for feat in to_ohe:
        dummies = pd.get_dummies(data=data_t[feat], prefix=feat, prefix_sep=
'_',
                                 drop_first=True)
        data_t.drop(feat, axis=1, inplace=True)
        data_t = pd.concat([data_t, dummies], axis=1)
    data_t.rename(lambda col: col.replace('.0', '').replace(' ','__')
                  .replace('^','_pow_'),
                  axis=1, inplace=True)
    return data_t
```

bin_latlong

Calculates (or looks up if it's test data) 5 equal divisions in each of lat and long. Then, for each row, it uses floor division by the bin size to calculate which bin the lat-long belongs in. The bin number increases by latitude first and then by longitude.

In [7]:
```
def bin_latlong(data_t, test=False):

    if not test:
        max_lat = data_t['lat'].max()
        min_lat = data_t['lat'].min()
        x=(max_lat - min_lat)/5

        max_long = data_t['long'].max()
        min_long = data_t['long'].min()
        y=(max_long - min_long)/5
        #record stats used:
        latlong_vals = [x, min_lat, y, min_long]
        with open('latlong_vals.pickle', 'wb') as f:
            pickle.dump(latlong_vals, f)
    else:
        with open('latlong_vals.pickle', 'rb') as f:
            latlong_vals = pickle.load(f)
        x = latlong_vals[0]
        min_lat = latlong_vals[1]
        y = latlong_vals[2]
        min_long = latlong_vals[3]


    data_t.loc[:,'lat_long'] = data_t.apply(lambda row: (row.lat-min_lat)//x
                                        + 5*(row.long-min_long)//y, axis=1)



    return data_t
```

bin_basement

Sets basement to 1 if it has a basement, otherwise it remains 0

In [8]:
```
def bin_basement(data_t):
    data_t.sqft_basement = data_t.sqft_basement.apply(lambda x:
                                                      1 if not x else 0)
    return data_t
```

simple_selector

simple_selector starts by calling the model function, which returns a model using all the features in x_cols. Only the features in x_cols with a pvalue lower than or equal to alpha are kept and returned.

In [9]:
```
def simple_selector(data_s, x_cols, alpha=0.1):

    results = model(data_s, x_cols)
    pv = pd.DataFrame(results.pvalues).drop('Intercept')
    pv.rename(columns={0:'p_value'}, inplace=True)
    x_cols = list(pv[pv.p_value <= alpha].index)

    return x_cols
```

stepwise_selector

This function loops adding or removing features depending on their pvalue each iteration; breaking out of the loop when stable.

In [10]:
```
def stepwise_selector(data_s,
                      x_cols=[],
                      alpha=0.05,
                      verbose=False):
    """
    Perform a forward-backward feature selection
    based on p-value from statsmodels.api.OLS
    Arguments:
        X - pandas.DataFrame with candidate features
        y - list-like with the target
        x_cols - list of features to start with (column names of X)
        threshold_in - include a feature if its p-value < threshold_in
        threshold_out - exclude a feature if its p-value > threshold_out
        verbose - whether to print the sequence of inclusions and exclusions
    Returns: list of selected features
    Always set threshold_in < threshold_out to avoid infinite looping.
    See https://en.wikipedia.org/wiki/Stepwise_regression for the details
    """

    X = data_s[x_cols]
    y = data_s['price']
    threshold_in = alpha - 0.02
    threshold_out = alpha + 0.02
    included = list(x_cols)
    while True:
        print(len(included))
        changed=False
        # forward step
        excluded = list(set(X.columns)-set(included))
        new_pval = pd.Series(index=excluded)
        for new_column in excluded:
            results = sm.OLS(y, sm.add_constant
                             (pd.DataFrame(X[included+[new_column]]))).fit()
            new_pval[new_column] = results.pvalues[new_column]
        best_pval = new_pval.min()
        if best_pval < threshold_in:
            best_feature = new_pval.idxmin()
            included.append(best_feature)
            changed=True
            if verbose:
                print('Add  {:30} with p-value {:.6}'.format(best_feature, bes
t_pval))

        # backward step
        results = sm.OLS(y, sm.add_constant(pd.DataFrame(X[included]))).fit()
        # use all coefs except intercept
        pvalues = results.pvalues.iloc[1:]
        worst_pval = pvalues.max() # null if pvalues is empty
        if worst_pval > threshold_out:
            changed=True
            worst_feature = pvalues.idxmax()
            included.remove(worst_feature)
            if verbose:
                print('Drop {:30} with p-value {:.6}'.format(worst_feature,
                                                             worst_pval))
        if not changed:
```

```
            break

    return included
```

rfe_selector

Here features are selected using sklearn's recursive-feature-elimination function for a range of numbers-of-features. Models are created for each set of features and the adjusted r-squared is plotted against the number of features. Using this graph, and the list of results, you can choose how many features you want to select. Enter this number when asked and it will call this function again. But this time specifying the number of features to select and returning with the selected features.
Note: tried using mpld3 to show y-values when hovering over markers but couldn't get the graph to appear before asking for input.

In [11]:

```
def rfe_selector(data_s, x_cols, nfts=None):
    X = data_s[x_cols]
    y = data_s[outcome]
    linreg = LR()
    r2_adj = []

    #if the desired n_features_to_select value is known, go straight to select
ing with that, otherwise start at 1
    n = nfts or 1
    #don't select more than the number of features, or 80 at most
    n_max = min(len(x_cols), 80)
    r = range(n, n_max+1)
    while n <= n_max:

        selector = RFE(linreg, n_features_to_select=n)
        results = selector.fit(X, y)

        #Can I use a metric from sklearn here, then move the for loop into the
if below it?

        #add column name from list (x_cols) if RFE selected it
        new_cols = []
        for i in range(len(x_cols)):
            if results.support_[i]:
                new_cols.append(x_cols[i])

        #append score for graphing
        results = model(data_s, new_cols)
        r2_adj.append(results.rsquared_adj)
        if nfts:
            return new_cols
        n+=1
    #show scores and graph for choosing the number of features to select
    print(list(zip(r, r2_adj)))
    fig, ax = plt.subplots()
    lines = ax.plot(r, r2_adj, marker='o')
    ax.set(xlabel='Number of Features', ylabel='Adj. R^2',
           title = 'Select Number of Features');

    plt.show(block=False)



    print('Select number of features')
    choice = int(input())
    new_cols = rfe_selector(data_s, x_cols, nfts=choice)

    return new_cols
```

Multicoll_remove

This function removes a feature from each pair with high multicollinearity
This function takes in the data, x_cols and the threshold for removing a feature

- create a dataframe of the correlation of the features in x_cols
- transform this to get a list of pairs of features with high multicollinearity
- for each pair: check whether they are the same feature or if one of the features has already been listed to be removed; if so, continue to next pair
- otherwise add the feature with the higher p-value to a list
- remove the features in this list from x_cols

return x_cols

In [12]:
```
def multicoll_remove(data_mr, x_cols, multicollinearity_threshold):

    pvalues = model(data_mr, x_cols).pvalues
    corr = data_mr[x_cols].corr().abs().stack().reset_index().sort_values(0,
                                                            ascending = False)
    corr['pairs'] = list(zip(corr.level_0, corr.level_1))
    corr = corr.set_index('pairs').drop(['level_0', 'level_1'], axis=1)
    corr.columns = ['cc']
    corr = corr[corr.cc > multicollinearity_threshold]

    to_drop = []
    for f0, f1 in corr.index:
        if (f0 == f1) | any(feat in [f0, f1] for feat in to_drop):
            continue
        to_drop.append(pvalues.loc[[f0, f1]]
                       .sort_values(ascending=False).index[0])
    x_cols = list(set(x_cols) - set(to_drop))
    return x_cols
```

metrics

metrics creates two graphs:

- a qqplot
- a residuals plot for homoscedasticity

In [13]:
```
def metrics(data_m, results, x_cols):
#     X = data[x_cols]
#     y = data.price
#     to_pred = pd.concat([y,X], axis=1)
#     yhat = results.predict(to_pred)
#     rmse = RMSE(y, yhat)

#     linreg = LR()
#     linreg.fit(X,y)
#     cv  = np.mean(CVS(linreg, X, y, cv=5,  scoring='neg_mean_squared_erro
r'))
#     rmse = (-cv)**0.5

#     print(f'RMSE: {rmse}')
    print(f'Number of features: {len(x_cols)}')

    fig = plt.figure(figsize=(14,14))
    ax1 = fig.add_subplot(211)
    sm.graphics.qqplot(results.resid, dist=stats.norm, line='45',
                               fit=True, alpha=0.1, ax=ax1)
    ax1.set(title='QQPlot')
    ax1.set(xlim=(-6, 6), ylim=(-7.5, 7.5))

    ax2 = fig.add_subplot(212)
    plt.scatter(data_m.price, results.resid, alpha=0.5);
    plt.hlines(0, xmax=data_m.price.max(), xmin=data_m.price.min());
#     ax2.title.set_text('Residuals Plot for Homoscedasticity')
    ax2.set(xlabel='Fitted', ylabel='Residuals',
       title = 'Residuals Plot for Homoscedasticity');
```

model

A small function to create a model using statsmodels' OLS function

In [14]:
```
def model(data_m, x_cols):

    predictors = '+'.join(x_cols)
    f = outcome + '~' + predictors
    results = ols(formula=f, data=data_m).fit()
    return results
```

unlog

In [15]:
```
def unlog(y):
    y = y.map(lambda x: np.e**x)
    return y
```

Main

Load. Join and Split:

In [16]:
```
data = pd.read_csv('data/clean.csv')
df = pd.read_csv('data/distances.csv')
data = pd.concat([data, df],axis=1)


data = data[data.price < 1000000]
data.sort_index(axis=1, inplace=True)
data_train, data_test = tts(data, train_size=0.8, random_state=111)
data_train = data_train.reset_index().drop('index', axis=1)
data_test = data_test.reset_index().drop('index', axis=1)
```

Modelling

Initial model using only cleaned data:

In [17]:
```
outcome = 'price'
x_cols = data_train.drop([outcome], axis=1).columns

results = model(data_train, x_cols)

metrics(data_train, results, x_cols)
results.summary()
```

```
Number of features: 36
```

Out[17]:

OLS Regression Results

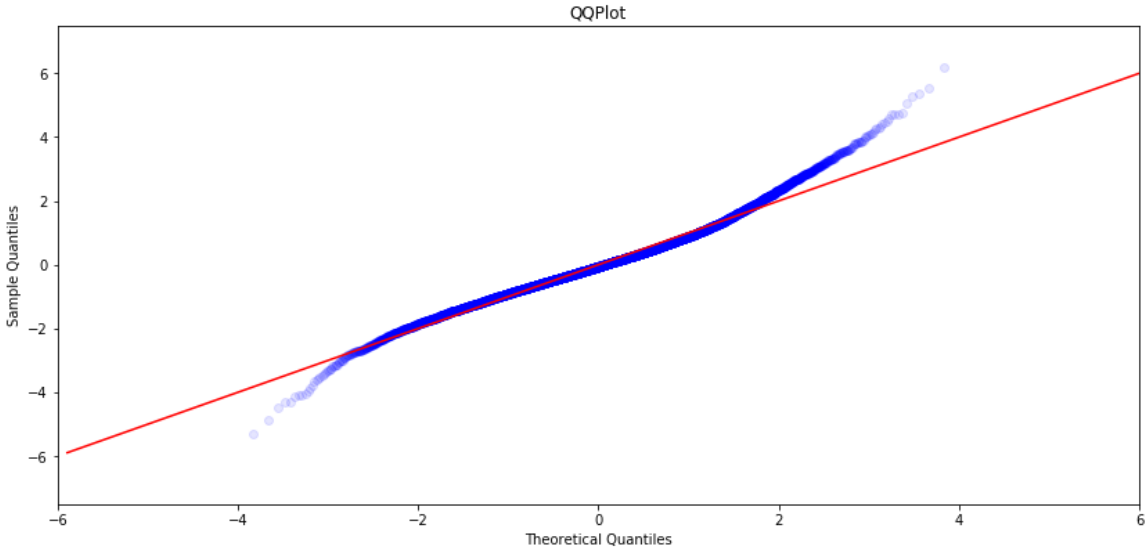
Dep. Variable:	price	**R-squared:**	0.785
Model:	OLS	**Adj. R-squared:**	0.785
Method:	Least Squares	**F-statistic:**	1604.
Date:	Tue, 01 Dec 2020	**Prob (F-statistic):**	0.00
Time:	10:51:04	**Log-Likelihood:**	-2.0305e+05
No. Observations:	15824	**AIC:**	4.062e+05
Df Residuals:	15787	**BIC:**	4.065e+05
Df Model:	36		
Covariance Type:	nonrobust		

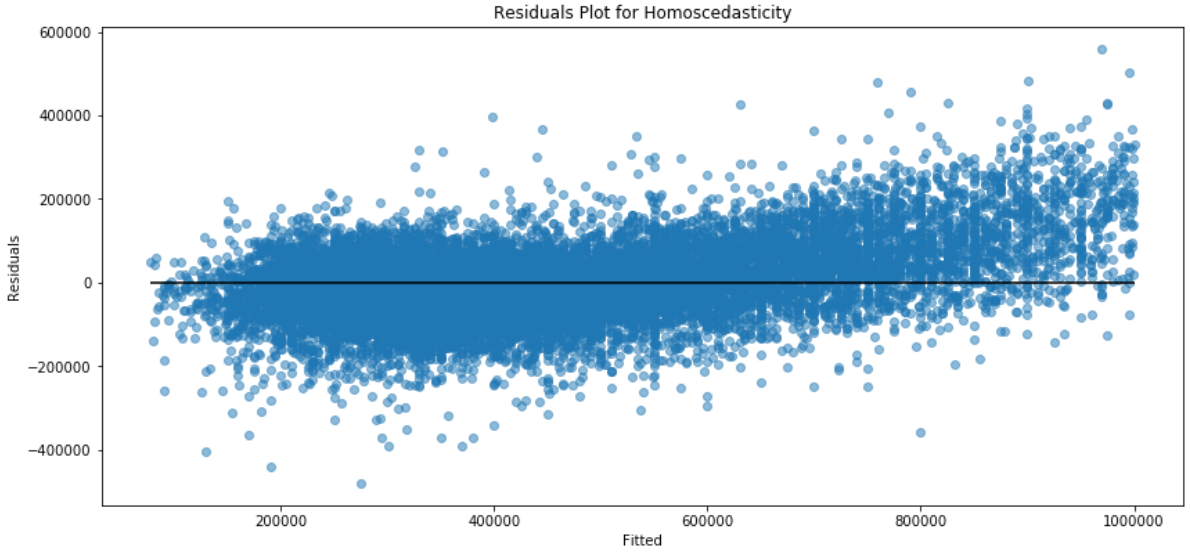
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.182e+08	4.87e+06	-24.292	0.000	-1.28e+08	-1.09e+08
Abandoned	-4131.0726	354.950	-11.638	0.000	-4826.815	-3435.331
Access_Point	-6363.7088	281.261	-22.626	0.000	-6915.013	-5812.404
Airport	-167.4957	307.199	-0.545	0.586	-769.641	434.650
Campground	3761.7398	249.134	15.099	0.000	3273.409	4250.071
Cemetery	-487.7177	513.632	-0.950	0.342	-1494.494	519.059
Commercial_Farm	-2887.6648	430.895	-6.702	0.000	-3732.269	-2043.061
Cultural	5824.9896	612.173	9.515	0.000	4625.061	7024.918
Educational	1.722e+04	1075.534	16.015	0.000	1.51e+04	1.93e+04
Fire	1879.2814	666.500	2.820	0.005	572.866	3185.697
Gate_wo_Building	-1.088e+04	542.686	-20.048	0.000	-1.19e+04	-9816.289
Gated_w_Building	9499.8348	254.467	37.332	0.000	9001.050	9998.619
Government	-1.029e+04	1056.108	-9.745	0.000	-1.24e+04	-8221.418
Lodging	4664.6963	635.221	7.343	0.000	3419.591	5909.801
Police	4011.2048	266.250	15.066	0.000	3489.325	4533.084
Public_Gathering	2702.3145	1474.465	1.833	0.067	-187.806	5592.435
Seasonal_Home	-4568.1764	361.342	-12.642	0.000	-5276.447	-3859.905
Utility	1842.0637	1608.313	1.145	0.252	-1310.413	4994.540
bathroomsx4	3579.3032	447.042	8.007	0.000	2703.050	4455.556
bedrooms	-5964.7190	1082.574	-5.510	0.000	-8086.688	-3842.750
condition	2.725e+04	1243.899	21.909	0.000	2.48e+04	2.97e+04
date	87.7155	6.412	13.681	0.000	75.148	100.283
floorsx2	-1208.8242	1036.973	-1.166	0.244	-3241.410	823.761
grade	4.632e+04	1225.348	37.801	0.000	4.39e+04	4.87e+04
lat	2.387e+05	2.13e+04	11.199	0.000	1.97e+05	2.8e+05

long	-8.852e+05	4.07e+04	-21.726	0.000	-9.65e+05	-8.05e+05
sqft_above	42.5919	10.320	4.127	0.000	22.363	62.821
sqft_basement	5.8718	10.212	0.575	0.565	-14.145	25.889
sqft_living	55.1441	10.337	5.335	0.000	34.882	75.406
sqft_living15	48.1394	2.078	23.163	0.000	44.066	52.213
sqft_lot	0.3178	0.038	8.384	0.000	0.244	0.392
sqft_lot15	-0.2487	0.105	-2.380	0.017	-0.454	-0.044
view	2.408e+04	1304.170	18.466	0.000	2.15e+04	2.66e+04
waterfront	1.682e+05	1.51e+04	11.174	0.000	1.39e+05	1.98e+05
yr_built	-1541.9461	73.074	-21.101	0.000	-1685.180	-1398.713
yr_renovated	763.8357	73.306	10.420	0.000	620.147	907.524
zipcode	2.8492	19.799	0.144	0.886	-35.958	41.657

Omnibus:	1068.572	**Durbin-Watson:**	2.014
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	2197.463
Skew:	0.463	**Prob(JB):**	0.00
Kurtosis:	4.574	**Cond. No.**	6.71e+08

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 6.71e+08. This might indicate that there are
strong multicollinearity or other numerical problems.


QQPlot
Sample Quantiles
Theoretical Quantiles
6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6


Residuals Plot for Homoscedasticity
Residuals
Fitted
600000
400000
200000
0
−200000
−400000
200000
400000
600000
800000
1000000

In [18]:
```
#simple feature selection
alpha=0.05
outcome = 'price'

data_t = data_train.copy()
x_cols = data_t.drop([outcome], axis=1).columns

x_cols = simple_selector(data_t, x_cols, alpha=alpha)

results = model(data_t, x_cols)
metrics(data_t, results, x_cols)
results.summary()
```

```
Number of features: 29
```

Out[18]:

OLS Regression Results

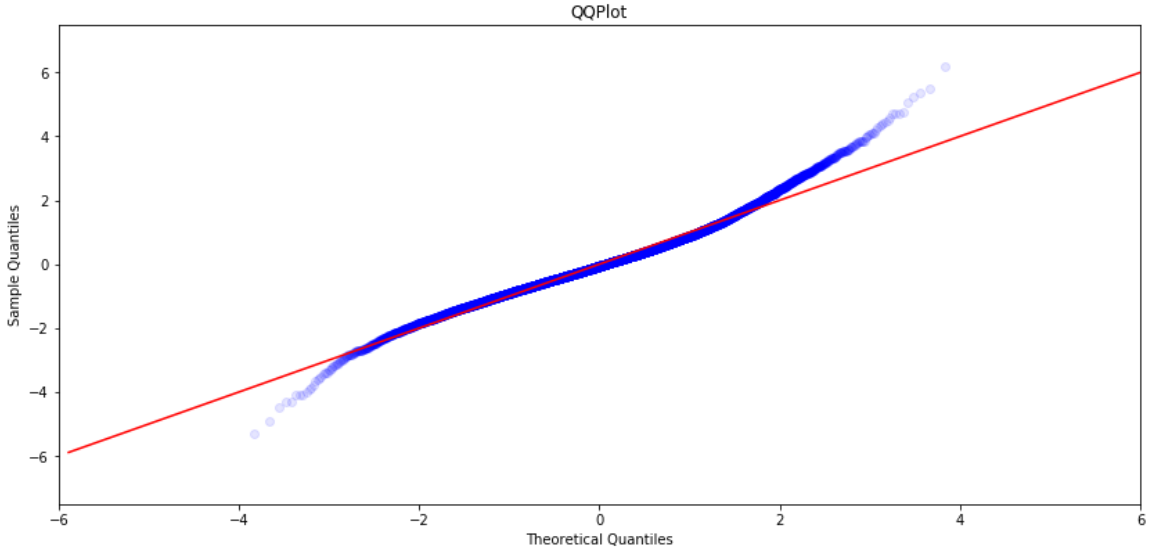
Dep. Variable:	price	**R-squared:**	0.785
Model:	OLS	**Adj. R-squared:**	0.785
Method:	Least Squares	**F-statistic:**	1991.
Date:	Tue, 01 Dec 2020	**Prob (F-statistic):**	0.00
Time:	10:51:07	**Log-Likelihood:**	-2.0306e+05
No. Observations:	15824	**AIC:**	4.062e+05
Df Residuals:	15794	**BIC:**	4.064e+05
Df Model:	29		
Covariance Type:	nonrobust		

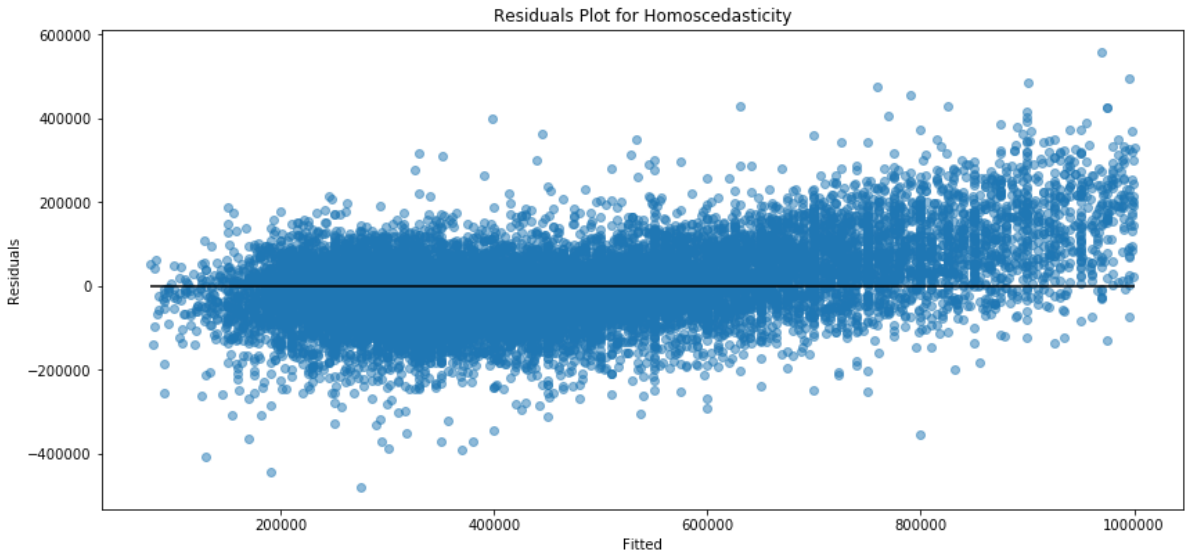
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.192e+08	3.27e+06	-36.442	0.000	-1.26e+08	-1.13e+08
Abandoned	-4205.6967	326.236	-12.892	0.000	-4845.157	-3566.236
Access_Point	-6400.4273	267.745	-23.905	0.000	-6925.237	-5875.617
Campground	3791.9113	214.273	17.697	0.000	3371.912	4211.911
Commercial_Farm	-2884.2634	426.114	-6.769	0.000	-3719.496	-2049.031
Cultural	5876.0593	602.683	9.750	0.000	4694.732	7057.387
Educational	1.754e+04	1016.699	17.257	0.000	1.56e+04	1.95e+04
Fire	2128.1176	645.532	3.297	0.001	862.802	3393.433
Gate_wo_Building	-1.092e+04	517.163	-21.121	0.000	-1.19e+04	-9909.418
Gated_w_Building	9326.3998	196.537	47.454	0.000	8941.166	9711.634
Government	-9663.2701	1005.400	-9.611	0.000	-1.16e+04	-7692.570
Lodging	4673.7518	585.631	7.981	0.000	3525.849	5821.655
Police	3938.3773	240.281	16.391	0.000	3467.398	4409.356
Seasonal_Home	-4626.2854	292.964	-15.791	0.000	-5200.528	-4052.043
bathroomsx4	3442.1117	434.061	7.930	0.000	2591.304	4292.920
bedrooms	-5952.6091	1081.097	-5.506	0.000	-8071.682	-3833.536
condition	2.737e+04	1237.457	22.120	0.000	2.49e+04	2.98e+04
date	87.6013	6.411	13.664	0.000	75.035	100.167
grade	4.624e+04	1223.329	37.800	0.000	4.38e+04	4.86e+04
lat	2.328e+05	2.04e+04	11.429	0.000	1.93e+05	2.73e+05
long	-8.982e+05	3.28e+04	-27.369	0.000	-9.62e+05	-8.34e+05
sqft_above	35.3458	2.226	15.881	0.000	30.983	39.709
sqft_living	61.8571	2.451	25.234	0.000	57.052	66.662
sqft_living15	48.3362	2.058	23.483	0.000	44.302	52.371
sqft_lot	0.3188	0.038	8.414	0.000	0.245	0.393

sqft_lot15	-0.1987	0.102	-1.942	0.052	-0.399	0.002
view	2.409e+04	1297.063	18.573	0.000	2.15e+04	2.66e+04
waterfront	1.685e+05	1.5e+04	11.229	0.000	1.39e+05	1.98e+05
yr_built	-1551.7235	72.482	-21.408	0.000	-1693.796	-1409.651
yr_renovated	758.4209	73.184	10.363	0.000	614.972	901.870

Omnibus:	1070.159	**Durbin-Watson:**	2.014
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	2195.329
Skew:	0.464	**Prob(JB):**	0.00
Kurtosis:	4.571	**Cond. No.**	1.27e+08

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.27e+08. This might indicate that there are strong multicollinearity or other numerical problems.





In [19]:
```
#simple feature selection with multicollinearity removal
multicollinearity_threshold=0.7
alpha=0.05
outcome = 'price'

data_t = data_train.copy()
x_cols = data_t.drop([outcome], axis=1).columns

x_cols = simple_selector(data_t, x_cols, alpha=alpha)
x_cols = multicoll_remove(data_t, x_cols, multicollinearity_threshold)

results = model(data_t, x_cols)
metrics(data_t, results, x_cols)
results.summary()
```

```
Number of features: 23
```

Out[19]:

OLS Regression Results

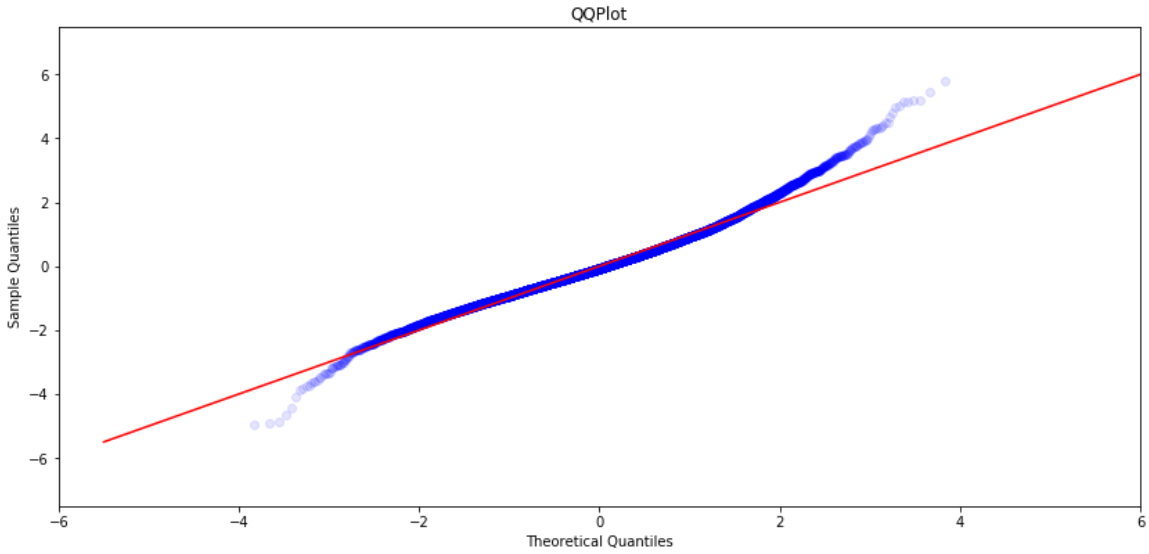
Dep. Variable:	price	**R-squared:**	0.756
Model:	OLS	**Adj. R-squared:**	0.755
Method:	Least Squares	**F-statistic:**	2127.
Date:	Tue, 01 Dec 2020	**Prob (F-statistic):**	0.00
Time:	10:51:10	**Log-Likelihood:**	-2.0407e+05
No. Observations:	15824	**AIC:**	4.082e+05
Df Residuals:	15800	**BIC:**	4.084e+05
Df Model:	23		
Covariance Type:	nonrobust		

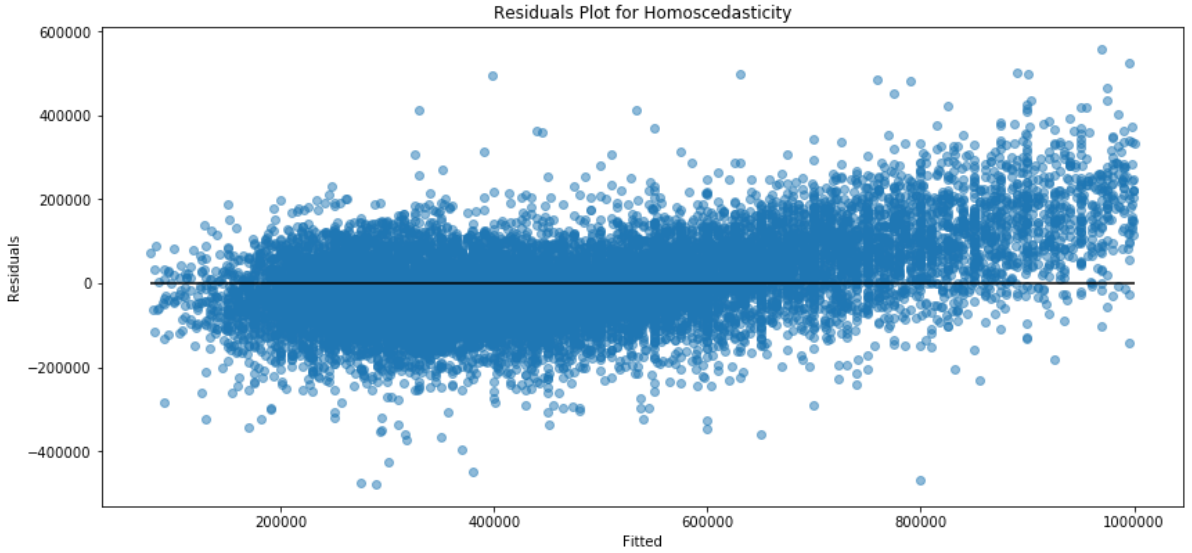
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-2.05e+07	1.29e+06	-15.839	0.000	-2.3e+07	-1.8e+07
sqft_lot	0.2360	0.040	5.855	0.000	0.157	0.315
Educational	-3172.9688	751.159	-4.224	0.000	-4645.326	-1700.612
Abandoned	-3969.6935	341.364	-11.629	0.000	-4638.807	-3300.580
long	9.518e+04	9391.553	10.135	0.000	7.68e+04	1.14e+05
lat	6.992e+05	7297.081	95.822	0.000	6.85e+05	7.14e+05
bedrooms	-6351.2985	1129.178	-5.625	0.000	-8564.616	-4137.981
Lodging	8638.9962	607.969	14.210	0.000	7447.308	9830.684
sqft_living	109.4069	1.792	61.038	0.000	105.894	112.920
Police	2891.8186	233.295	12.396	0.000	2434.534	3349.103
Campground	2078.6132	220.447	9.429	0.000	1646.512	2510.714
grade	6.442e+04	1217.761	52.904	0.000	6.2e+04	6.68e+04
waterfront	1.842e+05	1.6e+04	11.536	0.000	1.53e+05	2.15e+05
Government	-5103.8641	1063.992	-4.797	0.000	-7189.410	-3018.318
Gated_w_Building	1.032e+04	204.252	50.512	0.000	9916.726	1.07e+04
Commercial_Farm	-4446.9171	452.198	-9.834	0.000	-5333.277	-3560.557
sqft_lot15	0.0464	0.108	0.428	0.668	-0.166	0.259
yr_built	-768.2531	37.061	-20.729	0.000	-840.897	-695.609
Fire	-161.0795	677.678	-0.238	0.812	-1489.405	1167.246
Cultural	1.007e+04	629.808	15.993	0.000	8837.951	1.13e+04
condition	2.358e+04	1285.127	18.345	0.000	2.11e+04	2.61e+04
date	82.7368	6.829	12.116	0.000	69.351	96.122
view	2.443e+04	1358.362	17.985	0.000	2.18e+04	2.71e+04
Access_Point	-7170.4618	270.379	-26.520	0.000	-7700.435	-6640.489

Omnibus:	1047.866	**Durbin-Watson:**	2.033

Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1980.260
Skew:	0.480	**Prob(JB):**	0.00
Kurtosis:	4.443	**Cond. No.**	4.71e+07

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.71e+07. This might indicate that there are
strong multicollinearity or other numerical problems.





In [20]:

```
multicollinearity_threshold=0.7
alpha=0.05
outcome = 'price'

data_t = data_train.copy()
x_cols = data_t.drop([outcome], axis=1).columns


x_cols = multicoll_remove(data_t, x_cols, multicollinearity_threshold)
x_cols = simple_selector(data_t, x_cols)

results = model(data_t, x_cols)
metrics(data_t, results, x_cols)
results.summary()
```

```
Number of features: 28
```

Out[20]:

OLS Regression Results

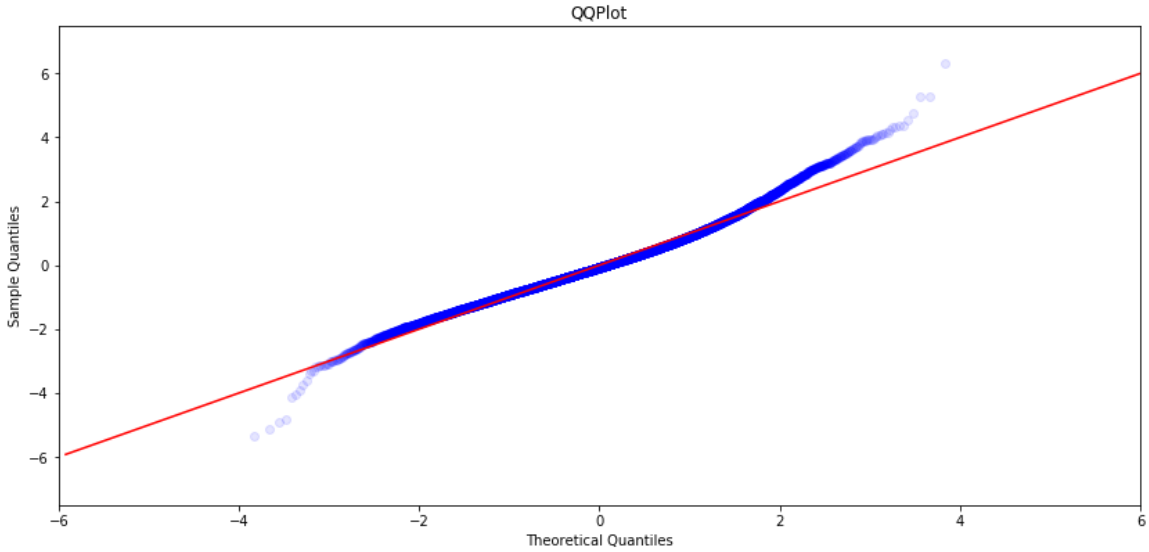
Dep. Variable:	price	**R-squared:**	0.749
Model:	OLS	**Adj. R-squared:**	0.749
Method:	Least Squares	**F-statistic:**	1686.
Date:	Tue, 01 Dec 2020	**Prob (F-statistic):**	0.00
Time:	10:51:13	**Log-Likelihood:**	-2.0428e+05
No. Observations:	15824	**AIC:**	4.086e+05
Df Residuals:	15795	**BIC:**	4.088e+05
Df Model:	28		
Covariance Type:	nonrobust		

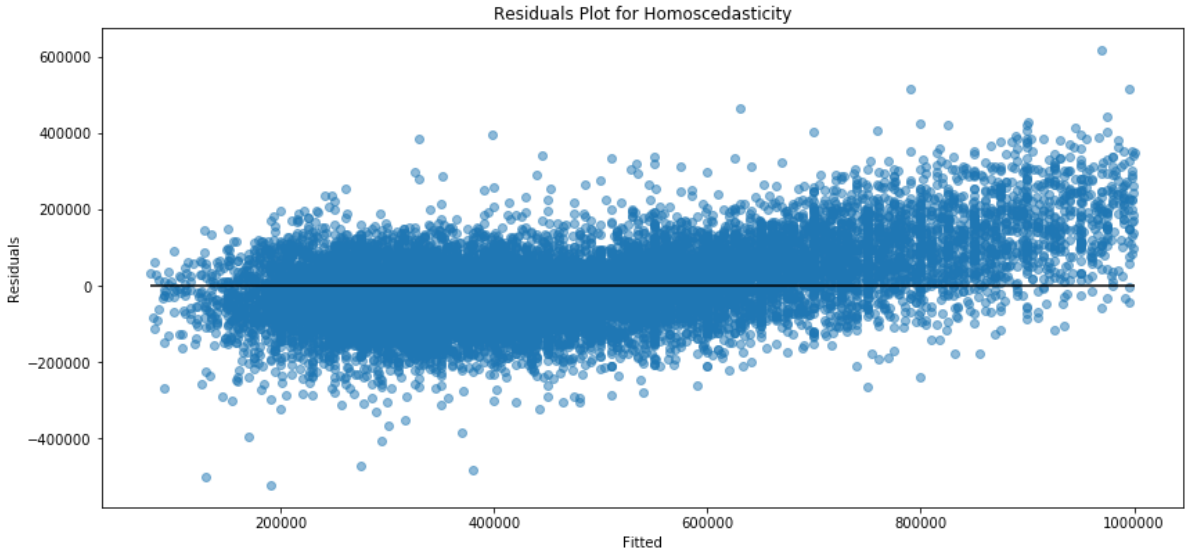
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-7.167e+06	2.07e+06	-3.460	0.001	-1.12e+07	-3.11e+06
sqft_lot	0.3611	0.035	10.212	0.000	0.292	0.430
long	6.489e+04	1.01e+04	6.395	0.000	4.5e+04	8.48e+04
lat	6.301e+05	9000.393	70.008	0.000	6.12e+05	6.48e+05
sqft_basement	13.3155	2.470	5.391	0.000	8.474	18.157
Lodging	6593.3667	656.691	10.040	0.000	5306.177	7880.557
Campground	3168.6897	261.670	12.109	0.000	2655.786	3681.594
Access_Point	-6464.7172	288.697	-22.393	0.000	-7030.596	-5898.839
grade	6.812e+04	1229.631	55.402	0.000	6.57e+04	7.05e+04
Gated_w_Building	8454.0197	241.492	35.007	0.000	7980.667	8927.372
Airport	2678.0814	302.692	8.848	0.000	2084.770	3271.392
Commercial_Farm	-3821.9780	459.519	-8.317	0.000	-4722.688	-2921.268
yr_built	-1261.7360	42.685	-29.560	0.000	-1345.403	-1178.069
view	2.341e+04	1408.227	16.626	0.000	2.07e+04	2.62e+04
Educational	-1583.4235	795.970	-1.989	0.047	-3143.616	-23.231
Abandoned	-4508.7842	364.638	-12.365	0.000	-5223.516	-3794.053
bedrooms	9733.6332	1088.292	8.944	0.000	7600.456	1.19e+04
floorsx2	1.03e+04	1079.293	9.544	0.000	8185.304	1.24e+04
Police	2191.9310	254.540	8.611	0.000	1693.004	2690.858
waterfront	1.932e+05	1.62e+04	11.910	0.000	1.61e+05	2.25e+05
Government	-7922.6148	1119.206	-7.079	0.000	-1.01e+04	-5728.843
sqft_living15	88.6635	2.006	44.204	0.000	84.732	92.595
Fire	-1428.4286	693.205	-2.061	0.039	-2787.189	-69.669
bathroomsx4	9436.4842	458.346	20.588	0.000	8538.074	1.03e+04
Cultural	6895.7193	644.013	10.707	0.000	5633.380	8158.058

Public_Gathering	5339.8054	1530.975	3.488	0.000	2338.919	8340.692
condition	2.513e+04	1325.414	18.958	0.000	2.25e+04	2.77e+04
zipcode	-131.8522	19.958	-6.606	0.000	-170.973	-92.731
date	85.7698	6.923	12.389	0.000	72.200	99.340

Omnibus:	1052.101	**Durbin-Watson:**	2.015
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1830.450
Skew:	0.507	**Prob(JB):**	0.00
Kurtosis:	4.322	**Cond. No.**	2.63e+08

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.63e+08. This might indicate that there are strong multicollinearity or other numerical problems.





Running the simple_selector before multicoll_remove is definitely performing better -removes more features and has a higher adj. R^2. Though running multicoll_remove either way is dropping the R^2, it should be making it more reliable.

What to Transform?

One-Hot-Encoding

Candidates for OHE:

- bedrooms
- bathroomsx4
- floorsx2
- view
- condition
- grade
- lat_long

Next, for each of the features above, I one-hot-encode it, run it through simple feature selection, model it, and compare the results. It needs to have a significant impact if I'm to transform it in the final model as it has to be worth the increase in columns. Note, only shows results if adj. R^2 is higher than the basemodel.

In [21]:
```
#Base model to compare against
multicollinearity_threshold=0.7
alpha=0.1
outcome = 'price'

data_t = data_train.copy()
x_cols = data_t.drop([outcome], axis=1).columns
x_cols = multicoll_remove(data_t, x_cols, multicollinearity_threshold)
x_cols = simple_selector(data_t, x_cols)

results = model(data_t, x_cols)
metrics(data_t, results, x_cols)
results.summary()
```

```
Number of features: 28
```

Out[21]:

OLS Regression Results

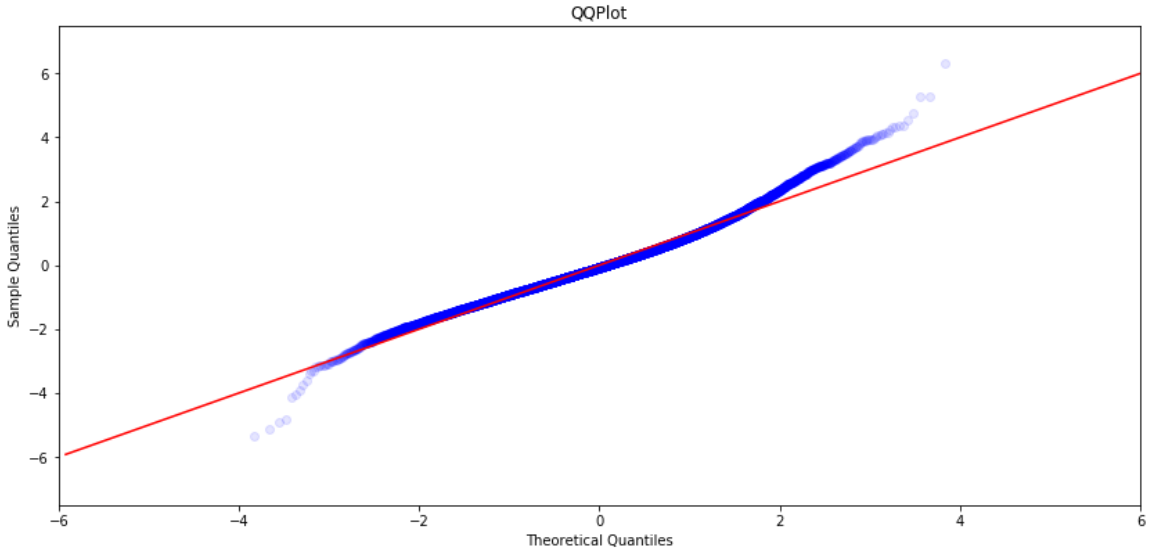
Dep. Variable:	price	**R-squared:**	0.749
Model:	OLS	**Adj. R-squared:**	0.749
Method:	Least Squares	**F-statistic:**	1686.
Date:	Tue, 01 Dec 2020	**Prob (F-statistic):**	0.00
Time:	10:51:16	**Log-Likelihood:**	-2.0428e+05
No. Observations:	15824	**AIC:**	4.086e+05
Df Residuals:	15795	**BIC:**	4.088e+05
Df Model:	28		
Covariance Type:	nonrobust		

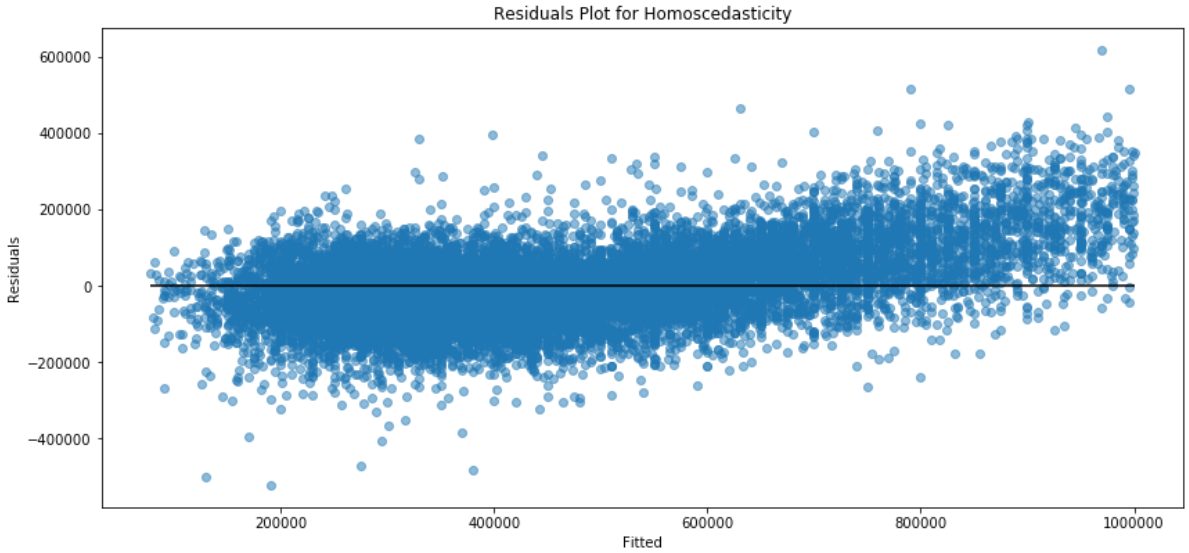
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-7.167e+06	2.07e+06	-3.460	0.001	-1.12e+07	-3.11e+06
sqft_lot	0.3611	0.035	10.212	0.000	0.292	0.430
long	6.489e+04	1.01e+04	6.395	0.000	4.5e+04	8.48e+04
lat	6.301e+05	9000.393	70.008	0.000	6.12e+05	6.48e+05
sqft_basement	13.3155	2.470	5.391	0.000	8.474	18.157
Lodging	6593.3667	656.691	10.040	0.000	5306.177	7880.557
Campground	3168.6897	261.670	12.109	0.000	2655.786	3681.594
Access_Point	-6464.7172	288.697	-22.393	0.000	-7030.596	-5898.839
grade	6.812e+04	1229.631	55.402	0.000	6.57e+04	7.05e+04
Gated_w_Building	8454.0197	241.492	35.007	0.000	7980.667	8927.372
Airport	2678.0814	302.692	8.848	0.000	2084.770	3271.392
Commercial_Farm	-3821.9780	459.519	-8.317	0.000	-4722.688	-2921.268
yr_built	-1261.7360	42.685	-29.560	0.000	-1345.403	-1178.069
view	2.341e+04	1408.227	16.626	0.000	2.07e+04	2.62e+04
Educational	-1583.4235	795.970	-1.989	0.047	-3143.616	-23.231
Abandoned	-4508.7842	364.638	-12.365	0.000	-5223.516	-3794.053
bedrooms	9733.6332	1088.292	8.944	0.000	7600.456	1.19e+04
floorsx2	1.03e+04	1079.293	9.544	0.000	8185.304	1.24e+04
Police	2191.9310	254.540	8.611	0.000	1693.004	2690.858
waterfront	1.932e+05	1.62e+04	11.910	0.000	1.61e+05	2.25e+05
Government	-7922.6148	1119.206	-7.079	0.000	-1.01e+04	-5728.843
sqft_living15	88.6635	2.006	44.204	0.000	84.732	92.595
Fire	-1428.4286	693.205	-2.061	0.039	-2787.189	-69.669
bathroomsx4	9436.4842	458.346	20.588	0.000	8538.074	1.03e+04
Cultural	6895.7193	644.013	10.707	0.000	5633.380	8158.058

Public_Gathering	5339.8054	1530.975	3.488	0.000	2338.919	8340.692
condition	2.513e+04	1325.414	18.958	0.000	2.25e+04	2.77e+04
zipcode	-131.8522	19.958	-6.606	0.000	-170.973	-92.731
date	85.7698	6.923	12.389	0.000	72.200	99.340

Omnibus:	1052.101	**Durbin-Watson:**	2.015
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1830.450
Skew:	0.507	**Prob(JB):**	0.00
Kurtosis:	4.322	**Cond. No.**	2.63e+08

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.63e+08. This might indicate that there are strong multicollinearity or other numerical problems.





In [22]:
```
#Run models to ohe each ohe candidate separately:
multicollinearity_threshold=0.7
alpha=0.1
cand_ohe = ['bedrooms', 'bathroomsx4', 'floorsx2', 'view', 'condition', 'grad
e',
            'lat_long']

for feat in cand_ohe:
    data_t = data_train.copy()
    if feat == 'lat_long':
        data_t = bin_latlong(data_t)
    to_ohe = [feat]
    data_t = ohe(data_t, to_ohe)
    x_cols = data_t.drop([outcome], axis=1).columns
    x_cols = multicoll_remove(data_t, x_cols, multicollinearity_threshold)
    x_cols = simple_selector(data_t, x_cols)


    results = model(data_t, x_cols)
    if results.rsquared_adj > 0.749:
        metrics(data_t, results, x_cols)
        print(feat)
        print(results.summary())
```

Number of features: 33
bedrooms

OLS Regression Results

Dep. Variable:	price	R-squared:	0.750
Model:	OLS	Adj. R-squared:	0.749
Method:	Least Squares	F-statistic:	1434.
Date:	Tue, 01 Dec 2020	Prob (F-statistic):	0.00
Time:	10:51:19	Log-Likelihood:	-2.0427e+05
No. Observations:	15824	AIC:	4.086e+05
Df Residuals:	15790	BIC:	4.089e+05
Df Model:	33		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-7e+06	2.07e+06	-3.379	0.001	-1.11e+07	-2.94e+06
sqft_lot	0.3640	0.035	10.303	0.000	0.295	0.433
long	6.517e+04	1.01e+04	6.427	0.000	4.53e+04	8.5e+04
lat	6.307e+05	8995.222	70.118	0.000	6.13e+05	6.48e+05
bedrooms_3	1.445e+04	7244.758	1.995	0.046	250.419	2.87e+04
sqft_basement	13.3185	2.475	5.382	0.000	8.468	18.169
Lodging	6548.3763	656.231	9.979	0.000	5262.088	7834.664
bedrooms_8	6.262e+04	4.07e+04	1.538	0.124	-1.72e+04	1.42e+05
Campground	3178.1791	261.598	12.149	0.000	2665.417	3690.941
Access_Point	-6512.7912	288.457	-22.578	0.000	-7078.200	-5947.383
grade	6.778e+04	1237.099	54.787	0.000	6.54e+04	7.02e+04
Gated_w_Building	8473.9963	241.311	35.116	0.000	8000.999	8946.994
Airport	2700.3464	302.634	8.923	0.000	2107.150	3293.543
Commercial_Farm	-3825.9096	459.237	-8.331	0.000	-4726.067	-2925.753
yr_built	-1260.4701	42.726	-29.501	0.000	-1344.219	-1176.721

```
view              2.361e+04   1407.381     16.779      0.000    2.09e+04
2.64e+04
Educational      -1559.3063    795.894     -1.959      0.050   -3119.350
0.737
Abandoned        -4508.0567    364.649    -12.363      0.000   -5222.810   -3
793.303
bedrooms_4        3.301e+04   7383.229      4.471      0.000    1.85e+04
4.75e+04
floorsx2          9956.3409   1080.205      9.217      0.000    7839.016
1.21e+04
Police            2160.6607    254.309      8.496      0.000    1662.187    2
659.135
bedrooms_6        2.579e+04   1.06e+04      2.438      0.015    5052.127
4.65e+04
waterfront        1.931e+05   1.62e+04     11.911      0.000    1.61e+05
2.25e+05
Government       -8021.4074   1118.605     -7.171      0.000   -1.02e+04   -5
828.815
bedrooms_2        9577.4592   7394.023      1.295      0.195   -4915.670
2.41e+04
sqft_living15       87.4289      2.023     43.212      0.000      83.463
91.395
Fire             -1425.6483    692.736     -2.058      0.040   -2783.489
-67.807
bathroomsx4       9723.8296    454.783     21.381      0.000    8832.404
1.06e+04
Cultural          6903.7409    644.049     10.719      0.000    5641.332    8
166.150
Public_Gathering  5435.6674   1529.774      3.553      0.000    2437.136    8
434.199
condition         2.501e+04   1327.818     18.836      0.000    2.24e+04
2.76e+04
zipcode           -133.3727     19.962     -6.681      0.000    -172.500
-94.245
bedrooms_5        3.199e+04   7937.240      4.030      0.000    1.64e+04
4.75e+04
date                85.8594      6.917     12.412      0.000      72.301
99.418
==============================================================================
=
Omnibus:                     1047.958   Durbin-Watson:                   2.01
5
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1810.76
7
Skew:                           0.507   Prob(JB):                         0.0
0
Kurtosis:                       4.310   Cond. No.                     2.63e+0
8
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
[2] The condition number is large, 2.63e+08. This might indicate that there a
re
strong multicollinearity or other numerical problems.
```

Number of features: 30
floorsx2

OLS Regression Results

Dep. Variable:	price	R-squared:	0.751
Model:	OLS	Adj. R-squared:	0.751
Method:	Least Squares	F-statistic:	1590.
Date:	Tue, 01 Dec 2020	Prob (F-statistic):	0.00
Time:	10:51:24	Log-Likelihood:	-2.0422e+05
No. Observations:	15824	AIC:	4.085e+05
Df Residuals:	15793	BIC:	4.087e+05
Df Model:	30		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-8.39e+06	2.06e+06	-4.067	0.000	-1.24e+07	-4.35e+06
sqft_lot	0.3773	0.035	10.697	0.000	0.308	0.446
long	5.579e+04	1.01e+04	5.503	0.000	3.59e+04	7.57e+04
lat	6.331e+05	8963.671	70.625	0.000	6.15e+05	6.51e+05
sqft_basement	17.3211	2.420	7.158	0.000	12.578	22.064
Lodging	6403.7648	653.795	9.795	0.000	5122.251	7685.278
floorsx2_5	6.714e+04	1.12e+04	5.974	0.000	4.51e+04	8.92e+04
Campground	3335.7117	256.955	12.982	0.000	2832.051	3839.372
Access_Point	-6785.6224	288.712	-23.503	0.000	-7351.531	-6219.714
grade	6.746e+04	1223.562	55.137	0.000	6.51e+04	6.99e+04
Gated_w_Building	8361.7692	240.649	34.747	0.000	7890.070	8833.468
Airport	2934.2693	302.111	9.713	0.000	2342.098	3526.441
Commercial_Farm	-3925.0374	457.660	-8.576	0.000	-4822.102	-3027.972
floorsx2_3	2.36e+04	3115.270	7.575	0.000	1.75e+04	2.97e+04
yr_built	-1140.3293	43.540	-26.190	0.000	-1225.673	-1054.986

```
floorsx2_4        3.081e+04   2352.875     13.094      0.000    2.62e+04
3.54e+04
view               2.406e+04   1403.248     17.143      0.000    2.13e+04
2.68e+04
Educational       -1487.3053    793.760     -1.874      0.061   -3043.165
68.554
Abandoned         -4475.1472    361.418    -12.382      0.000   -5183.568   -3
766.726
bedrooms           8455.1221   1097.862      7.701      0.000    6303.187
1.06e+04
Police             2137.8821    253.568      8.431      0.000    1640.859    2
634.905
waterfront         1.932e+05   1.62e+04     11.961      0.000    1.62e+05
2.25e+05
Government        -8160.2771   1114.501     -7.322      0.000   -1.03e+04   -5
975.728
sqft_living15        86.0707      2.011     42.796      0.000      82.129
90.013
Fire              -1543.8278    690.216     -2.237      0.025   -2896.730    -
190.926
bathroomsx4        8854.8865    454.546     19.481      0.000    7963.925    9
745.848
Cultural           6895.6151    641.714     10.746      0.000    5637.782    8
153.449
Public_Gathering   5896.9478   1525.604      3.865      0.000    2906.591    8
887.305
condition          2.569e+04   1321.609     19.441      0.000    2.31e+04
2.83e+04
zipcode            -134.2557     19.875     -6.755      0.000    -173.213
-95.298
date                 86.5465      6.896     12.550      0.000      73.030
100.063
==============================================================================
=
Omnibus:                     1092.191   Durbin-Watson:                   2.01
8
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1946.48
1
Skew:                           0.515   Prob(JB):                         0.0
0
Kurtosis:                       4.376   Cond. No.                     2.63e+0
8
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
[2] The condition number is large, 2.63e+08. This might indicate that there a
re
strong multicollinearity or other numerical problems.
Number of features: 31
view
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.75
```

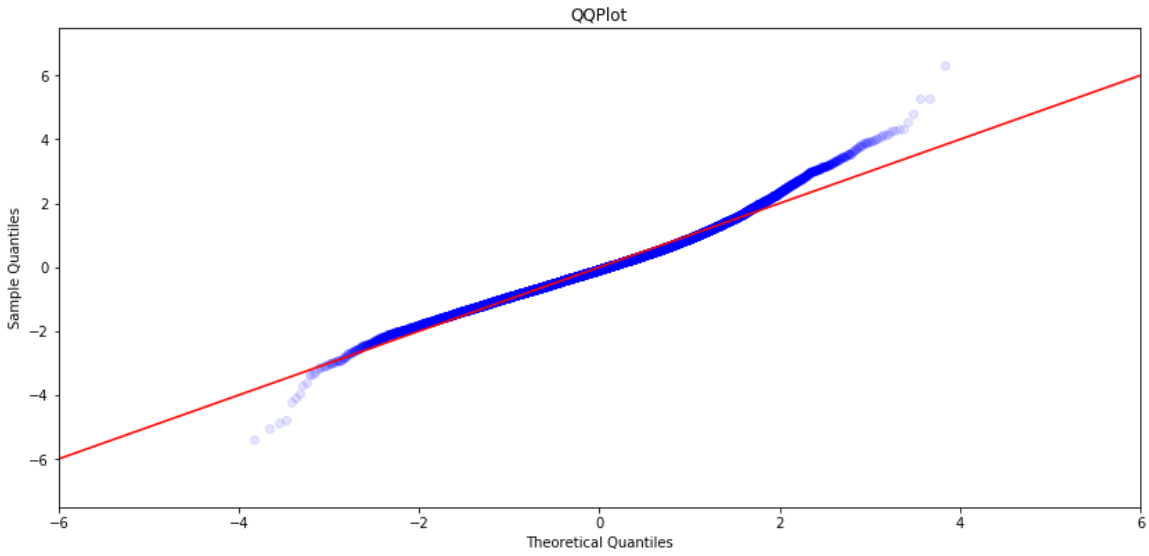
```
0
Model:                            OLS   Adj. R-squared:                  0.75
0
Method:                 Least Squares   F-statistic:                     152
9.
Date:                Tue, 01 Dec 2020   Prob (F-statistic):               0.0
0
Time:                        10:51:25   Log-Likelihood:            -2.0426e+0
5
No. Observations:               15824   AIC:                         4.086e+0
5
Df Residuals:                   15792   BIC:                         4.088e+0
5
Df Model:                          31
Covariance Type:            nonrobust
==============================================================================
=======
                       coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------------------
-------
Intercept        -6.665e+06   2.07e+06     -3.218      0.001   -1.07e+07   -
2.61e+06
sqft_lot             0.3660      0.035     10.364      0.000       0.297
0.435
long                6.75e+04   1.01e+04      6.651      0.000    4.76e+04
8.74e+04
lat                 6.29e+05   8988.964     69.979      0.000    6.11e+05
6.47e+05
sqft_basement       12.8094      2.468      5.190      0.000       7.972
17.647
Lodging           6474.0950    655.982      9.869      0.000    5188.294    7
759.896
view_2            3.957e+04   4242.615      9.327      0.000    3.13e+04
4.79e+04
Campground        3229.4071    261.546     12.347      0.000    2716.747    3
742.068
Access_Point     -6447.7668    288.274    -22.367      0.000   -7012.816   -5
882.717
grade             6.822e+04   1227.903     55.561      0.000    6.58e+04
7.06e+04
Gated_w_Building  8414.0791    241.337     34.864      0.000    7941.030    8
887.128
Airport           2687.8058    302.307      8.891      0.000    2095.250    3
280.362
Commercial_Farm  -3866.8094    458.890     -8.426      0.000   -4766.286   -2
967.333
yr_built         -1258.3322     42.629    -29.518      0.000   -1341.890   -1
174.775
view_1             6.28e+04   6843.212      9.177      0.000    4.94e+04
7.62e+04
view_3            5.896e+04   6437.842      9.159      0.000    4.63e+04
7.16e+04
Educational      -1563.5385    794.879     -1.967      0.049   -3121.593
-5.484
Abandoned        -4565.7616    364.240    -12.535      0.000   -5279.713   -3
851.810
```

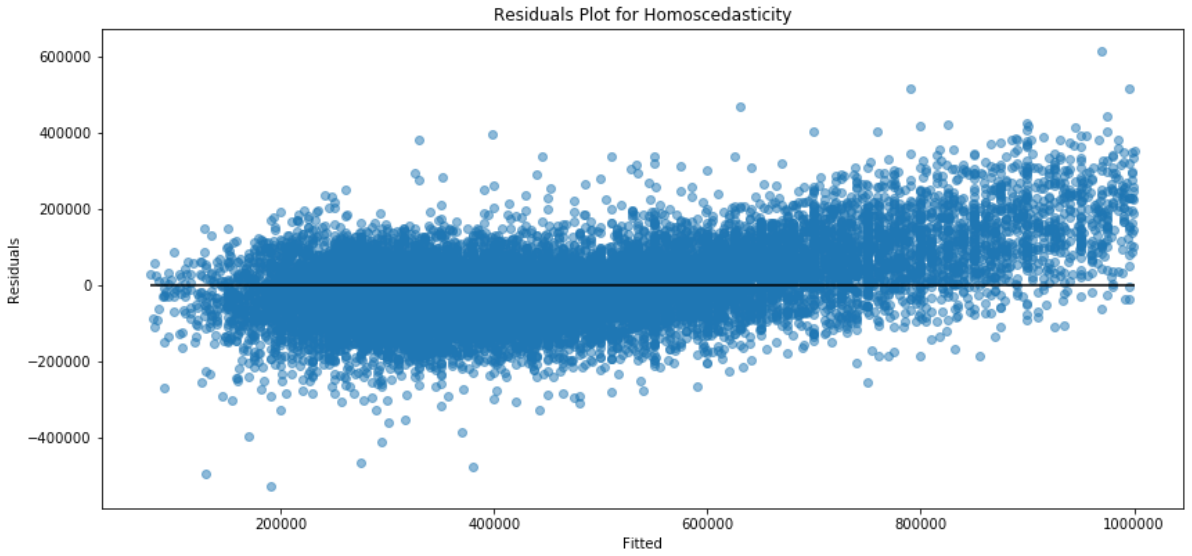
```
bedrooms          9865.0309   1086.853      9.077      0.000    7734.674
1.2e+04
floorsx2          1.019e+04   1078.142      9.452      0.000    8077.122
1.23e+04
Police            2195.9669    254.214      8.638      0.000    1697.679    2
694.255
waterfront        1.692e+05   1.79e+04      9.454      0.000    1.34e+05
2.04e+05
Government       -7909.3406   1117.714     -7.076      0.000   -1.01e+04   -5
718.494
sqft_living15       88.0849      2.005     43.929      0.000      84.155
92.015
Fire             -1426.9062    692.180     -2.061      0.039   -2783.657
-70.155
view_4            1.287e+05   1.15e+04     11.188      0.000    1.06e+05
1.51e+05
bathroomsx4       9475.0199    457.782     20.698      0.000    8577.714
1.04e+04
Cultural          6887.1831    643.051     10.710      0.000    5626.730    8
147.636
Public_Gathering  5379.7087   1528.693      3.519      0.000    2383.296    8
376.121
condition         2.511e+04   1323.492     18.974      0.000    2.25e+04
2.77e+04
zipcode           -133.2702     19.934     -6.686      0.000    -172.343
-94.198
date                85.7089      6.913     12.399      0.000      72.159
99.259
==============================================================================
=
Omnibus:                     1052.091   Durbin-Watson:                   2.01
4
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1850.56
5
Skew:                           0.503   Prob(JB):                         0.0
0
Kurtosis:                       4.339   Cond. No.                     2.63e+0
8
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
[2] The condition number is large, 2.63e+08. This might indicate that there a
re
strong multicollinearity or other numerical problems.
Number of features: 48
lat_long
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.76
8
Model:                            OLS   Adj. R-squared:                  0.76
7
Method:                 Least Squares   F-statistic:                     108
```

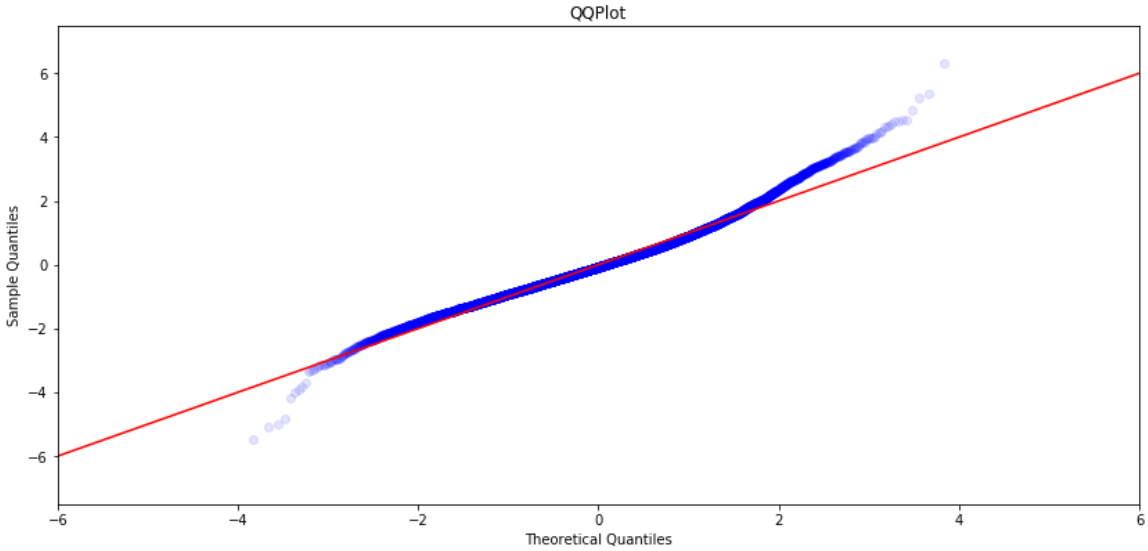
```
7.
Date:                Tue, 01 Dec 2020   Prob (F-statistic):               0.0
0
Time:                        10:51:31   Log-Likelihood:            -2.0367e+0
5
No. Observations:               15824   AIC:                         4.074e+0
5
Df Residuals:                   15775   BIC:                         4.078e+0
5
Df Model:                          48
Covariance Type:            nonrobust
==============================================================================
=======
                       coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------------------
-------
Intercept        -2.336e+07   2.12e+06    -11.019      0.000   -2.75e+07   -
1.92e+07
sqft_lot             0.4142      0.034     12.161      0.000       0.347
0.481
lat_long_12        -6.9e+04   1.25e+04     -5.523      0.000   -9.35e+04   -
4.45e+04
lat_long_28      -1.316e+04   3.77e+04     -0.349      0.727   -8.71e+04
6.08e+04
lat                4.22e+05   1.34e+04     31.514      0.000    3.96e+05
4.48e+05
sqft_basement       15.5463      2.380      6.532      0.000      10.881
20.211
lat_long_15      -1.834e+05   1.69e+04    -10.848      0.000   -2.17e+05    -
1.5e+05
Lodging           1841.5882    689.402      2.671      0.008     490.282    3
192.894
Campground        6027.6623    293.562     20.533      0.000    5452.247    6
603.077
Access_Point     -6596.5998    326.661    -20.194      0.000   -7236.892   -5
956.308
grade             6.382e+04   1198.035     53.275      0.000    6.15e+04
6.62e+04
Gated_w_Building  8879.1496    295.088     30.090      0.000    8300.744    9
457.556
lat_long_5       -2.163e+04   5187.187     -4.170      0.000   -3.18e+04   -
1.15e+04
Airport           2178.9252    329.719      6.608      0.000    1532.638    2
825.212
Commercial_Farm  -3612.8942    483.617     -7.471      0.000   -4560.840   -2
664.949
lat_long_20       -3.82e+05   9.66e+04     -3.954      0.000   -5.71e+05   -
1.93e+05
yr_built         -1258.1658     41.341    -30.434      0.000   -1339.198   -1
177.134
lat_long_24      -2.097e+05   9.93e+04     -2.112      0.035   -4.04e+05   -
1.51e+04
lat_long_6       -4.617e+04   6307.922     -7.319      0.000   -5.85e+04   -
3.38e+04
lat_long_14      -9916.6250   1.56e+04     -0.636      0.525   -4.05e+04
2.07e+04
```

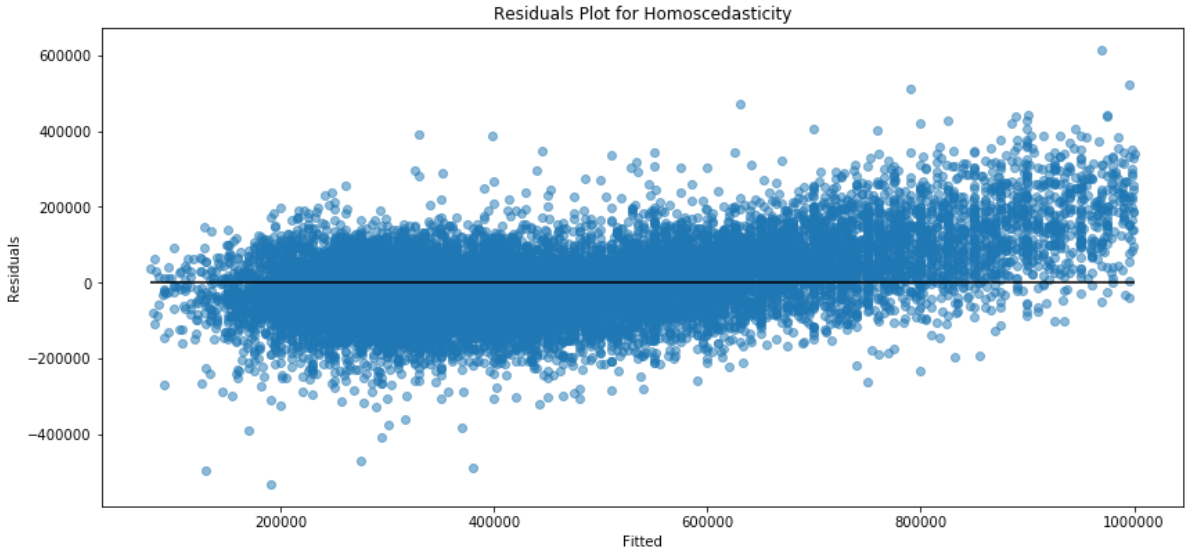
view	2.397e+04	1361.544	17.602	0.000	2.13e+04	2.66e+04
Abandoned	-4668.0057	377.858	-12.354	0.000	-5408.651	-3927.360
lat_long_27	-6.623e+04	4.49e+04	-1.475	0.140	-1.54e+05	2.18e+04
lat_long_17	-2.625e+05	2.03e+04	-12.899	0.000	-3.02e+05	-2.23e+05
lat_long_8	-5.29e+04	8081.441	-6.545	0.000	-6.87e+04	-3.71e+04
bedrooms	1.061e+04	1050.631	10.099	0.000	8550.821	1.27e+04
lat_long_3	-6.344e+04	8263.944	-7.677	0.000	-7.96e+04	-4.72e+04
lat_long_19	-2.681e+05	4.61e+04	-5.820	0.000	-3.58e+05	-1.78e+05
floorsx2	1.153e+04	1045.969	11.026	0.000	9482.256	1.36e+04
Police	3199.3230	325.988	9.814	0.000	2560.349	3838.297
lat_long_25	-2.419e+05	6.94e+04	-3.486	0.000	-3.78e+05	-1.06e+05
Utility	1.057e+04	1714.734	6.167	0.000	7213.033	1.39e+04
waterfront	1.936e+05	1.57e+04	12.306	0.000	1.63e+05	2.24e+05
lat_long_9	-3.785e+04	9216.131	-4.107	0.000	-5.59e+04	-1.98e+04
lat_long_10	-5.658e+04	9955.637	-5.683	0.000	-7.61e+04	-3.71e+04
Government	-5069.2373	1090.745	-4.647	0.000	-7207.223	-2931.252
Gate_wo_Building	-6163.4509	450.912	-13.669	0.000	-7047.290	-5279.612
lat_long_11	-6.501e+04	1.1e+04	-5.887	0.000	-8.67e+04	-4.34e+04
sqft_living15	81.9805	1.956	41.914	0.000	78.147	85.814
lat_long_18	-2.624e+05	2.4e+04	-10.915	0.000	-3.1e+05	-2.15e+05
Fire	2747.2862	710.045	3.869	0.000	1355.516	4139.056
lat_long_16	-1.903e+05	1.71e+04	-11.097	0.000	-2.24e+05	-1.57e+05
bathroomsx4	9519.6338	441.936	21.541	0.000	8653.388	1.04e+04
Cultural	4630.2302	709.584	6.525	0.000	3239.364	6021.097
lat_long_13	-6.786e+04	1.39e+04	-4.885	0.000	-9.51e+04	-4.06e+04
condition	2.279e+04	1285.789	17.724	0.000	2.03e+04	2.53e+04
zipcode	55.5879	22.988	2.418	0.016	10.529	100.646
lat_long_7	-8.003e+04	7162.561	-11.173	0.000	-9.41e+04	-6.6e+04
date	85.4834	6.669	12.818	0.000	72.411	

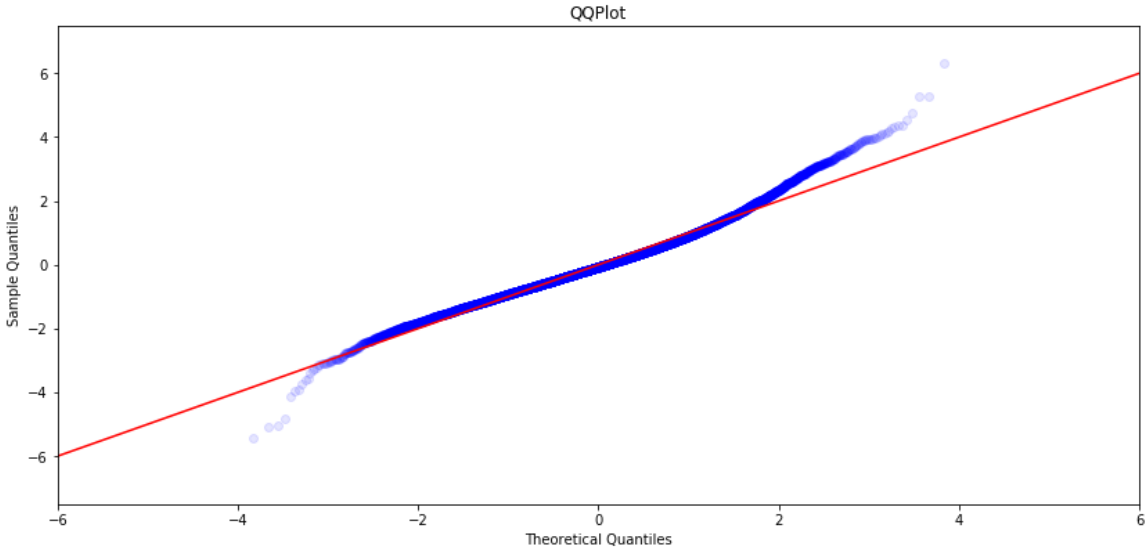
```
98.556
==============================================================================
=
Omnibus:                     1170.408   Durbin-Watson:                   2.01
3
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2283.22
4
Skew:                           0.517   Prob(JB):                         0.0
0
Kurtosis:                       4.548   Cond. No.                     2.80e+0
8
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
[2] The condition number is large, 2.8e+08. This might indicate that there ar
e
strong multicollinearity or other numerical problems.
```

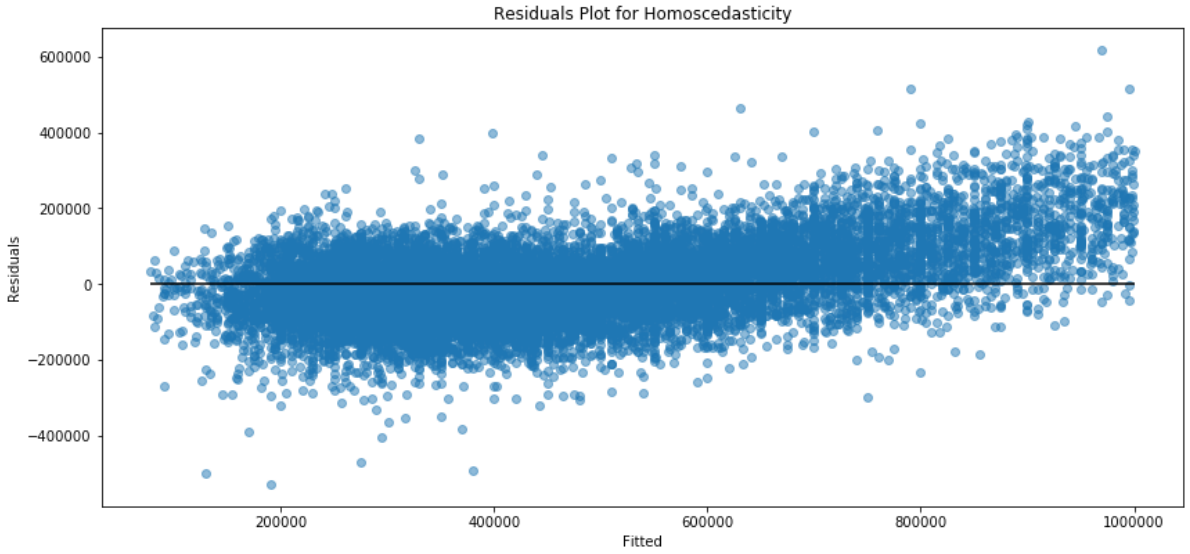


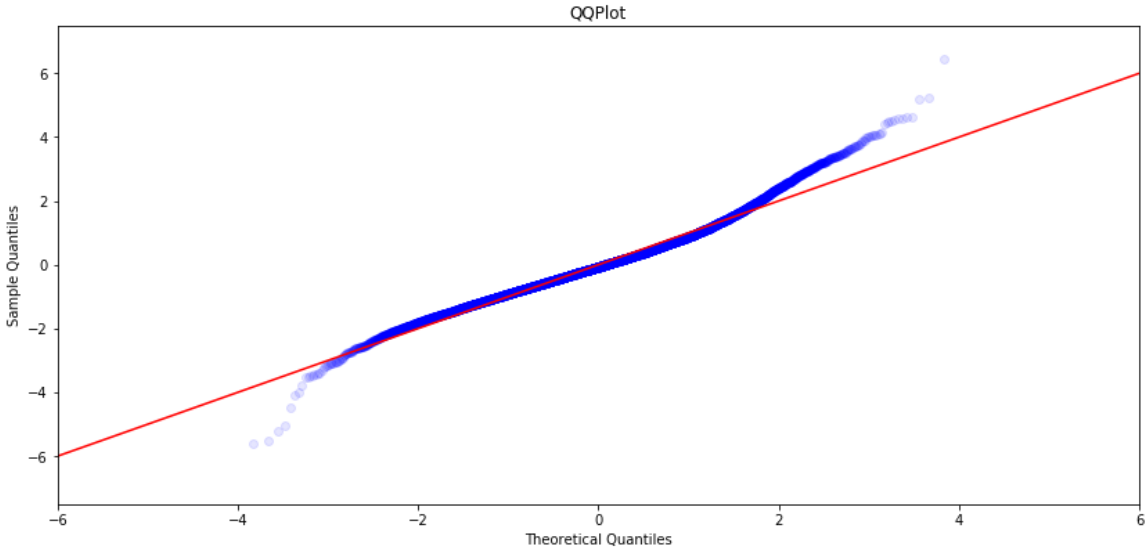


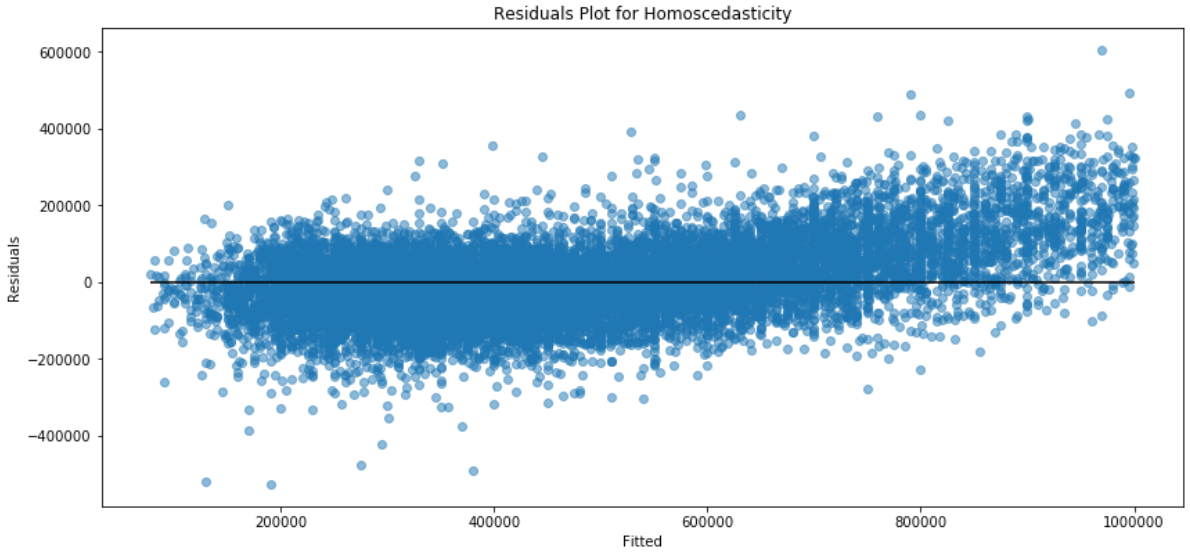

QQPlot
Sample Quantiles
Theoretical Quantiles
6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6


Residuals Plot for Homoscedasticity
Residuals
Fitted
600000
400000
200000
0
−200000
−400000
200000
400000
600000
800000
1000000


QQPlot
Sample Quantiles
Theoretical Quantiles
6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6


Residuals Plot for Homoscedasticity
Residuals
Fitted
600000
400000
200000
0
−200000
−400000
200000
400000
600000
800000
1000000





The only OHE candidate that made any significant improvement was the binned lat_long feature

Log

As above but for log transforms. Note, the score increase only needs to be large enough to make the decrease in interpretability worthwhile.

In [23]:

```
data_t = data_train.copy()
multicollinearity_threshold=0.7
alpha=0.1
cand_log = data_t.drop(['lat', 'long', 'waterfront', 'zipcode'], axis=1).colum
ns

for feat in cand_log:
    data_t = data_train.copy()
    to_log = [feat]
    data_t = log(data_t, to_log)
    x_cols = data_t.drop([outcome], axis=1).columns
    x_cols = multicoll_remove(data_t, x_cols, multicollinearity_threshold)
    x_cols = simple_selector(data_t, x_cols)


    results = model(data_t, x_cols)
    if results.rsquared_adj > 0.749:
        metrics(data_t, results, x_cols)
        print(feat)
        print(results.summary())
```

Number of features: 26
Gate_wo_Building

OLS Regression Results

Dep. Variable:	price	R-squared:	0.753
Model:	OLS	Adj. R-squared:	0.753
Method:	Least Squares	F-statistic:	1853.
Date:	Tue, 01 Dec 2020	Prob (F-statistic):	0.00
Time:	10:51:46	Log-Likelihood:	-2.0416e+05
No. Observations:	15824	AIC:	4.084e+05
Df Residuals:	15797	BIC:	4.086e+05
Df Model:	26		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-2.746e+07	4.55e+05	-60.376	0.000	-2.84e+07	-2.66e+07
sqft_lot	0.3673	0.035	10.476	0.000	0.299	0.436
lat	6.2e+05	9195.710	67.426	0.000	6.02e+05	6.38e+05
sqft_basement	14.9321	2.450	6.095	0.000	10.130	19.735
Lodging	7348.0420	654.644	11.224	0.000	6064.865	8631.219
Campground	2994.9030	250.107	11.974	0.000	2504.665	3485.141
Access_Point	-5624.6184	279.829	-20.100	0.000	-6173.115	-5076.122
grade	6.749e+04	1218.612	55.381	0.000	6.51e+04	6.99e+04
Gated_w_Building	9412.0339	254.690	36.955	0.000	8912.813	9911.254
Airport	1380.0970	307.297	4.491	0.000	777.759	1982.435
Commercial_Farm	-3324.4850	427.511	-7.776	0.000	-4162.455	-2486.515
yr_built	-1270.1994	42.158	-30.129	0.000	-1352.835	-1187.564
Cemetery	-3933.8107	391.296	-10.053	0.000	-4700.795	-3166.826
view	2.371e+04	1391.685	17.034	0.000	2.1e+04	2.64e+04
Abandoned	-3857.7460	356.775	-10.813	0.000	-4557.065	-3158.427

```
bedrooms          1.002e+04   1079.341      9.281      0.000    7902.246
1.21e+04
floorsx2          1.018e+04   1068.396      9.524      0.000    8081.599
1.23e+04
Police            3744.5786    271.293     13.803      0.000    3212.814    4
276.343
waterfront        1.919e+05   1.61e+04     11.940      0.000     1.6e+05
2.23e+05
Government       -8213.8497   1109.208     -7.405      0.000   -1.04e+04   -6
039.676
sqft_living15       86.4772      1.988     43.491      0.000      82.580
90.375
bathroomsx4       9409.6581    454.597     20.699      0.000    8518.595
1.03e+04
Cultural          5706.1449    643.786      8.863      0.000    4444.252    6
968.038
Seasonal_Home    -2873.2085    149.809    -19.179      0.000   -3166.852   -2
579.565
Public_Gathering  6343.8583   1459.112      4.348      0.000    3483.833    9
203.884
condition          2.46e+04   1310.796     18.764      0.000     2.2e+04
2.72e+04
date                 85.0915      6.869     12.388      0.000      71.627
98.556
==============================================================================
=
Omnibus:                     1142.766   Durbin-Watson:                   2.01
3
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2116.95
1
Skew:                           0.522   Prob(JB):                         0.0
0
Kurtosis:                       4.456   Cond. No.                     1.54e+0
7
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
[2] The condition number is large, 1.54e+07. This might indicate that there a
re
strong multicollinearity or other numerical problems.
Number of features: 27
Police
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.75
4
Model:                            OLS   Adj. R-squared:                  0.75
4
Method:                 Least Squares   F-statistic:                     179
8.
Date:                Tue, 01 Dec 2020   Prob (F-statistic):               0.0
0
Time:                        10:51:48   Log-Likelihood:           -2.0412e+0
```

```
5
No. Observations:               15824   AIC:                         4.083e+0
5
Df Residuals:                   15796   BIC:                         4.085e+0
5
Df Model:                          27
Covariance Type:            nonrobust
================================================================================
=======
                       coef    std err          t      P>|t|      [0.025
0.975]
--------------------------------------------------------------------------------
-------
Intercept        -2.287e+07   5.54e+05    -41.275      0.000    -2.4e+07   -
2.18e+07
sqft_lot             0.3580      0.035     10.235      0.000       0.289
0.427
lat                5.23e+05   1.13e+04     46.411      0.000    5.01e+05
5.45e+05
sqft_basement       13.3283      2.438      5.466      0.000       8.549
18.108
Lodging           7358.3254    647.397     11.366      0.000    6089.353    8
627.297
Campground        3154.5105    256.987     12.275      0.000    2650.786    3
658.235
Access_Point     -5940.8029    273.043    -21.758      0.000   -6475.998   -5
405.608
grade             6.772e+04   1216.235     55.676      0.000    6.53e+04
7.01e+04
Gated_w_Building  1.039e+04    261.707     39.684      0.000    9872.725
1.09e+04
Airport           1032.4253    309.892      3.332      0.001     425.001    1
639.849
Commercial_Farm  -3360.0454    447.815     -7.503      0.000   -4237.813   -2
482.277
yr_built         -1257.5390     41.976    -29.959      0.000   -1339.817   -1
175.261
Cemetery         -5851.7251    437.688    -13.370      0.000   -6709.643   -4
993.807
view              2.337e+04   1386.809     16.852      0.000    2.07e+04
2.61e+04
Educational       3667.4903    789.350      4.646      0.000    2120.274    5
214.706
Abandoned        -3094.7301    352.270     -8.785      0.000   -3785.219   -2
404.241
bedrooms          9886.4254   1076.058      9.188      0.000    7777.228
1.2e+04
floorsx2          1.008e+04   1066.376      9.451      0.000    7988.447
1.22e+04
Police            2.248e+04   1497.358     15.014      0.000    1.95e+04
2.54e+04
waterfront        1.901e+05    1.6e+04     11.868      0.000    1.59e+05
2.22e+05
Government       -8030.8862   1105.540     -7.264      0.000   -1.02e+04   -5
863.901
Gate_wo_Building -3454.4280    171.448    -20.149      0.000   -3790.485   -3
118.371
```

```
sqft_living15       86.5587      1.982     43.678      0.000      82.674
90.443
bathroomsx4       9434.7396    453.355     20.811      0.000    8546.113
1.03e+04
Cultural          6154.8990    637.870      9.649      0.000    4904.602    7
405.196
Public_Gathering  3342.4494   1511.304      2.212      0.027     380.122    6
304.777
condition          2.43e+04   1307.366     18.589      0.000    2.17e+04
2.69e+04
date                85.1728      6.850     12.433      0.000      71.745
98.601
==============================================================================
=
Omnibus:                     1116.208   Durbin-Watson:                   2.01
4
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2083.58
0
Skew:                           0.510   Prob(JB):                         0.0
0
Kurtosis:                       4.456   Cond. No.                     1.88e+0
7
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
[2] The condition number is large, 1.88e+07. This might indicate that there a
re
strong multicollinearity or other numerical problems.
Number of features: 28
Public_Gathering
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.74
9
Model:                            OLS   Adj. R-squared:                  0.74
9
Method:                 Least Squares   F-statistic:                     168
8.
Date:                Tue, 01 Dec 2020   Prob (F-statistic):               0.0
0
Time:                        10:51:49   Log-Likelihood:            -2.0428e+0
5
No. Observations:               15824   AIC:                          4.086e+0
5
Df Residuals:                   15795   BIC:                          4.088e+0
5
Df Model:                          28
Covariance Type:            nonrobust
==============================================================================
=======
                       coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------------------
```

```
-------
Intercept          -7.2e+06   2.07e+06     -3.478      0.001   -1.13e+07   -
3.14e+06
sqft_lot             0.3609      0.035     10.248      0.000       0.292
0.430
long              6.626e+04   1.01e+04      6.546      0.000    4.64e+04
8.61e+04
lat               6.314e+05   9000.486     70.150      0.000    6.14e+05
6.49e+05
sqft_basement       13.3701      2.468      5.417      0.000       8.532
18.208
Lodging           6618.1782    651.982     10.151      0.000    5340.219    7
896.138
Campground        3188.5147    261.123     12.211      0.000    2676.684    3
700.345
Access_Point     -6472.7835    288.345    -22.448      0.000   -7037.973   -5
907.594
grade             6.802e+04   1229.499     55.319      0.000    6.56e+04
7.04e+04
Gated_w_Building  8551.9635    242.968     35.198      0.000    8075.718    9
028.209
Airport           2648.6990    302.686      8.751      0.000    2055.400    3
241.999
Commercial_Farm  -3842.5302    458.769     -8.376      0.000   -4741.771   -2
943.290
yr_built         -1266.0531     42.662    -29.676      0.000   -1349.676   -1
182.430
view              2.323e+04   1408.011     16.501      0.000    2.05e+04
2.6e+04
Educational      -1299.6776    773.530     -1.680      0.093   -2815.885
216.529
Abandoned        -4546.5930    363.976    -12.491      0.000   -5260.027   -3
833.159
bedrooms          9698.4854   1087.844      8.915      0.000    7566.187
1.18e+04
floorsx2          1.045e+04   1079.394      9.680      0.000    8332.799
1.26e+04
Police            2186.4731    254.048      8.607      0.000    1688.510    2
684.436
waterfront        1.929e+05   1.62e+04     11.895      0.000    1.61e+05
2.25e+05
Government       -7717.4534   1085.713     -7.108      0.000   -9845.575   -5
589.331
sqft_living15       88.3262      2.005     44.064      0.000      84.397
92.255
Fire             -1283.0863    688.173     -1.864      0.062   -2631.985
65.812
bathroomsx4       9455.2376    458.168     20.637      0.000    8557.177
1.04e+04
Cultural          6968.8173    642.810     10.841      0.000    5708.836    8
228.798
Public_Gathering  6131.7024   1235.252      4.964      0.000    3710.467    8
552.938
condition         2.504e+04   1325.119     18.896      0.000    2.24e+04
2.76e+04
zipcode           -130.2620     19.952     -6.529      0.000    -169.370
-91.154
```

```
date                85.8546      6.920     12.407      0.000      72.291
99.418
==============================================================================
=
Omnibus:                     1052.859   Durbin-Watson:                   2.01
5
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1837.58
9
Skew:                           0.506   Prob(JB):                         0.0
0
Kurtosis:                       4.327   Cond. No.                     2.63e+0
8
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
[2] The condition number is large, 2.63e+08. This might indicate that there a
re
strong multicollinearity or other numerical problems.
Number of features: 28
Seasonal_Home
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.75
3
Model:                            OLS   Adj. R-squared:                  0.75
3
Method:                 Least Squares   F-statistic:                     171
9.
Date:                Tue, 01 Dec 2020   Prob (F-statistic):               0.0
0
Time:                        10:51:49   Log-Likelihood:            -2.0417e+0
5
No. Observations:               15824   AIC:                             4.084e+0
5
Df Residuals:                   15795   BIC:                             4.086e+0
5
Df Model:                          28
Covariance Type:            nonrobust
==============================================================================
=======
                       coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------------------
-------
Intercept        -1.947e+07   1.94e+06    -10.009      0.000   -2.33e+07   -
1.57e+07
sqft_lot             0.3601      0.035     10.263      0.000       0.291
0.429
lat               5.298e+05   1.19e+04     44.414      0.000    5.06e+05
5.53e+05
sqft_basement       13.2714      2.448      5.422      0.000       8.473
18.069
Lodging           7285.7516    655.212     11.120      0.000    6001.462    8
```

```
570.042
Campground        3409.8446    260.179     13.106      0.000    2899.864    3
919.825
Access_Point     -6275.6979    280.531    -22.371      0.000   -6825.571   -5
725.824
grade             6.792e+04   1219.792     55.685      0.000    6.55e+04
7.03e+04
Gated_w_Building  9812.3538    263.705     37.210      0.000    9295.462
1.03e+04
Airport           1580.7705    308.529      5.124      0.000     976.019    2
185.522
Commercial_Farm  -3241.2191    450.394     -7.196      0.000   -4124.043   -2
358.395
yr_built         -1256.7328     42.343    -29.680      0.000   -1339.729   -1
173.736
Cemetery         -5557.2367    462.638    -12.012      0.000   -6464.060   -4
650.414
view              2.365e+04   1397.392     16.927      0.000    2.09e+04
2.64e+04
Educational       3361.5359    799.653      4.204      0.000    1794.124    4
928.948
Abandoned        -3414.5485    372.512     -9.166      0.000   -4144.715   -2
684.382
bedrooms          9934.7421   1080.365      9.196      0.000    7817.103
1.21e+04
floorsx2          9856.1471   1071.554      9.198      0.000    7755.780
1.2e+04
Police            3014.3116    267.640     11.263      0.000    2489.707    3
538.917
waterfront        1.908e+05   1.61e+04     11.861      0.000    1.59e+05
2.22e+05
Government       -7482.6255   1110.452     -6.738      0.000   -9659.237   -5
306.014
Gate_wo_Building -3054.3307    185.220    -16.490      0.000   -3417.383   -2
691.278
sqft_living15       87.3176      1.988     43.929      0.000      83.422
91.214
bathroomsx4       9419.3499    454.933     20.705      0.000    8527.630
1.03e+04
Cultural          6369.4055    639.744      9.956      0.000    5115.435    7
623.376
Public_Gathering  4430.6843   1512.406      2.930      0.003    1466.197    7
395.172
condition         2.471e+04   1316.117     18.772      0.000    2.21e+04
2.73e+04
zipcode            -37.9895     20.569     -1.847      0.065     -78.307
2.328
date                85.7596      6.872     12.480      0.000      72.290
99.229
==============================================================================
=
Omnibus:                     1137.198   Durbin-Watson:                   2.01
3
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2104.82
4
Skew:                           0.520   Prob(JB):                         0.0
0
```

```
Kurtosis:                       4.453   Cond. No.                     2.49e+0
8
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
[2] The condition number is large, 2.49e+08. This might indicate that there a
re
strong multicollinearity or other numerical problems.
Number of features: 28
floorsx2
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.75
0
Model:                            OLS   Adj. R-squared:                  0.74
9
Method:                 Least Squares   F-statistic:                     169
1.
Date:                Tue, 01 Dec 2020   Prob (F-statistic):               0.0
0
Time:                        10:51:52   Log-Likelihood:            -2.0427e+0
5
No. Observations:               15824   AIC:                         4.086e+0
5
Df Residuals:                   15795   BIC:                         4.088e+0
5
Df Model:                          28
Covariance Type:            nonrobust
==============================================================================
=======
                       coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------------------
-------
Intercept        -7.296e+06   2.07e+06     -3.527      0.000   -1.14e+07   -
3.24e+06
sqft_lot             0.3655      0.035     10.347      0.000       0.296
0.435
long              6.263e+04   1.01e+04      6.175      0.000    4.28e+04
8.25e+04
lat               6.304e+05   8989.507     70.125      0.000    6.13e+05
6.48e+05
sqft_basement       15.7157      2.489      6.313      0.000      10.836
20.595
Lodging           6590.1296    655.871     10.048      0.000    5304.548    7
875.711
Campground        3118.3434    260.820     11.956      0.000    2607.107    3
629.580
Access_Point     -6528.1960    288.524    -22.626      0.000   -7093.736   -5
962.656
grade             6.779e+04   1229.023     55.162      0.000    6.54e+04
7.02e+04
Gated_w_Building  8438.0177    241.235     34.978      0.000    7965.170    8
```

```
910.865
Airport           2724.3669    302.410      9.009      0.000    2131.609    3
317.125
Commercial_Farm  -3828.5793    458.988     -8.341      0.000   -4728.249   -2
928.910
yr_built         -1255.1335     42.066    -29.837      0.000   -1337.587   -1
172.680
view              2.352e+04   1406.645     16.722      0.000    2.08e+04
2.63e+04
Educational      -1625.3675    795.056     -2.044      0.041   -3183.769
-66.966
Abandoned        -4426.8207    364.193    -12.155      0.000   -5140.680   -3
712.961
bedrooms          9395.6767   1088.376      8.633      0.000    7262.336
1.15e+04
floorsx2          3.703e+04   3293.254     11.244      0.000    3.06e+04
4.35e+04
Police            2171.6496    254.273      8.541      0.000    1673.245    2
670.054
waterfront        1.924e+05   1.62e+04     11.872      0.000    1.61e+05
2.24e+05
Government       -7926.6793   1117.755     -7.092      0.000   -1.01e+04   -5
735.751
sqft_living15       88.1774      2.004     43.994      0.000      84.249
92.106
Fire             -1416.7696    692.383     -2.046      0.041   -2773.919
-59.620
bathroomsx4       9110.7150    459.990     19.806      0.000    8209.083
1e+04
Cultural          6891.3600    643.291     10.713      0.000    5630.436    8
152.284
Public_Gathering  5459.0802   1529.359      3.570      0.000    2461.362    8
456.798
condition         2.536e+04   1324.451     19.149      0.000    2.28e+04
2.8e+04
zipcode           -133.6305     19.938     -6.702      0.000    -172.710
-94.551
date                85.9833      6.915     12.434      0.000      72.428
99.538
==============================================================================
=
Omnibus:                     1060.355   Durbin-Watson:                   2.01
6
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1858.97
9
Skew:                           0.508   Prob(JB):                         0.0
0
Kurtosis:                       4.337   Cond. No.                     2.63e+0
8
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
[2] The condition number is large, 2.63e+08. This might indicate that there a
re
```

strong multicollinearity or other numerical problems.
Number of features: 28
price

OLS Regression Results

Dep. Variable:	price	R-squared:	0.768
Model:	OLS	Adj. R-squared:	0.767
Method:	Least Squares	F-statistic:	1865.
Date:	Tue, 01 Dec 2020	Prob (F-statistic):	0.00
Time:	10:51:53	Log-Likelihood:	2284.5
No. Observations:	15824	AIC:	-4511.
Df Residuals:	15795	BIC:	-4289.
Df Model:	28		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-34.7787	4.172	-8.337	0.000	-42.956	-26.602
sqft_lot	7.645e-07	7.52e-08	10.163	0.000	6.17e-07	9.12e-07
lat	1.3221	0.026	51.411	0.000	1.272	1.372
sqft_basement	4.209e-05	5.28e-06	7.977	0.000	3.17e-05	5.24e-05
Lodging	0.0201	0.001	14.024	0.000	0.017	0.023
Campground	0.0097	0.001	17.296	0.000	0.009	0.011
Access_Point	-0.0121	0.001	-19.722	0.000	-0.013	-0.011
grade	0.1387	0.003	52.744	0.000	0.134	0.144
Gated_w_Building	0.0212	0.001	37.016	0.000	0.020	0.022
Airport	0.0017	0.001	2.576	0.010	0.000	0.003
Commercial_Farm	-0.0055	0.001	-5.721	0.000	-0.007	-0.004
yr_built	-0.0021	9.14e-05	-22.933	0.000	-0.002	-0.002
Cemetery	-0.0120	0.001	-13.091	0.000	-0.014	-0.010
view	0.0467	0.003	15.502	0.000	0.041	0.053
Abandoned	-0.0080	0.001	-10.009	0.000	-0.010	

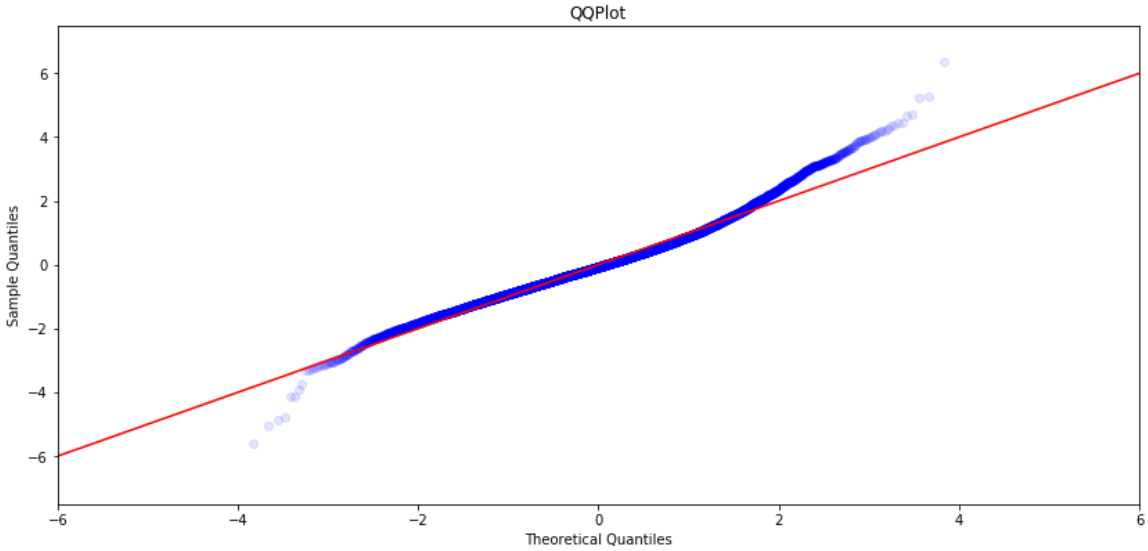
```
-0.006
bedrooms             0.0253      0.002     10.877      0.000       0.021
0.030
floorsx2             0.0221      0.002      9.589      0.000       0.018
0.027
Police               0.0093      0.001     16.087      0.000       0.008
0.010
Utility              0.0101      0.004      2.723      0.006       0.003
0.017
waterfront           0.4102      0.035     11.820      0.000       0.342
0.478
Government          -0.0116      0.002     -5.005      0.000      -0.016
-0.007
Gate_wo_Building    -0.0076      0.000    -19.057      0.000      -0.008
-0.007
sqft_living15        0.0002    4.3e-06     40.652      0.000       0.000
0.000
Fire                -0.0043      0.001     -2.920      0.004      -0.007
-0.001
bathroomsx4          0.0226      0.001     23.060      0.000       0.021
0.025
Cultural             0.0134      0.001      9.671      0.000       0.011
0.016
condition            0.0606      0.003     21.337      0.000       0.055
0.066
zipcode             -0.0001    4.4e-05     -3.022      0.003      -0.000   -
4.67e-05
date                 0.0002   1.48e-05     13.580      0.000       0.000
0.000
==============================================================================
=
Omnibus:                      711.815   Durbin-Watson:                   2.00
3
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2047.99
0
Skew:                          -0.190   Prob(JB):                         0.0
0
Kurtosis:                       4.721   Cond. No.                     2.47e+0
8
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
[2] The condition number is large, 2.47e+08. This might indicate that there a
re
strong multicollinearity or other numerical problems.
Number of features: 28
sqft_lot
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.75
1
Model:                            OLS   Adj. R-squared:                  0.75
1
```

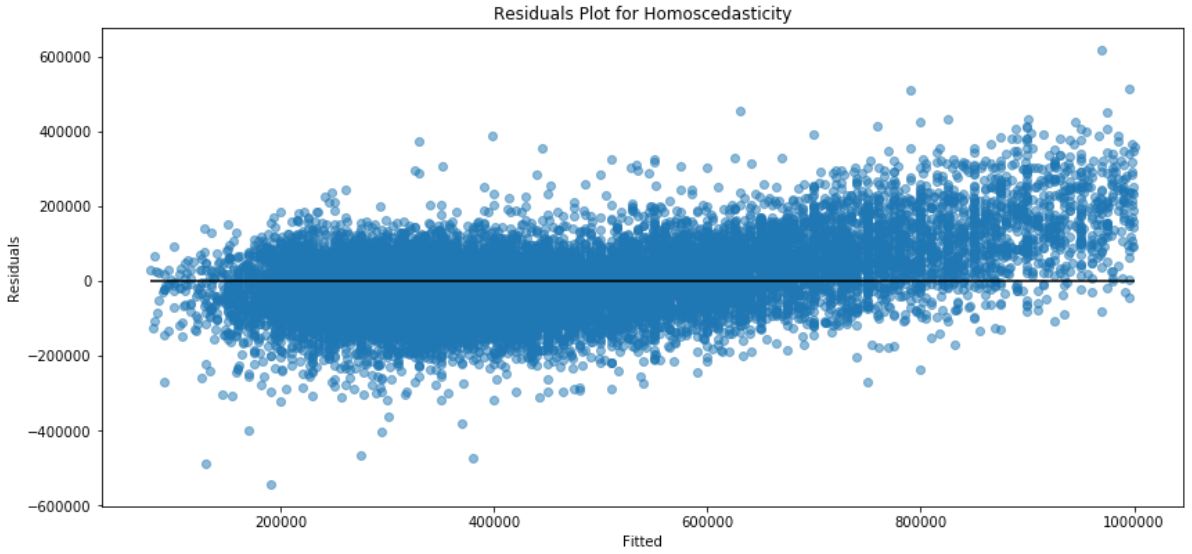
```
Method:                 Least Squares   F-statistic:                     170
3.
Date:                Tue, 01 Dec 2020   Prob (F-statistic):               0.0
0
Time:                        10:51:55   Log-Likelihood:            -2.0422e+0
5
No. Observations:               15824   AIC:                         4.085e+0
5
Df Residuals:                   15795   BIC:                         4.087e+0
5
Df Model:                          28
Covariance Type:            nonrobust
==============================================================================
=======
                       coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------------------
-------
Intercept        -8.264e+06   2.07e+06     -4.001      0.000   -1.23e+07   -
4.22e+06
sqft_lot          2.129e+04   1411.830     15.083      0.000    1.85e+04
2.41e+04
long              6.706e+04   1.01e+04      6.649      0.000    4.73e+04
8.68e+04
lat               6.377e+05   8988.657     70.942      0.000     6.2e+05
6.55e+05
sqft_basement       15.5273      2.462      6.306      0.000      10.701
20.354
Lodging           6448.0731    654.261      9.856      0.000    5165.647    7
730.499
Campground        3938.9629    265.980     14.809      0.000    3417.611    4
460.315
Access_Point     -6572.3843    287.714    -22.844      0.000   -7136.336   -6
008.433
grade             6.672e+04   1230.617     54.215      0.000    6.43e+04
6.91e+04
Gated_w_Building  8657.9224    241.089     35.912      0.000    8185.361    9
130.484
Airport           2723.0105    301.553      9.030      0.000    2131.932    3
314.089
Commercial_Farm  -3729.7024    457.748     -8.148      0.000   -4626.940   -2
832.465
yr_built         -1128.4330     43.897    -25.706      0.000   -1214.477   -1
042.389
view              2.304e+04   1403.134     16.423      0.000    2.03e+04
2.58e+04
Educational      -1828.4940    793.027     -2.306      0.021   -3382.917    -
274.071
Abandoned        -5435.7223    370.115    -14.687      0.000   -6161.189   -4
710.255
bedrooms          7633.1779   1091.628      6.992      0.000    5493.462    9
772.894
floorsx2          1.477e+04   1121.245     13.170      0.000    1.26e+04
1.7e+04
Police            2282.8799    253.524      9.005      0.000    1785.943    2
779.816
waterfront        1.886e+05   1.62e+04     11.670      0.000    1.57e+05
```

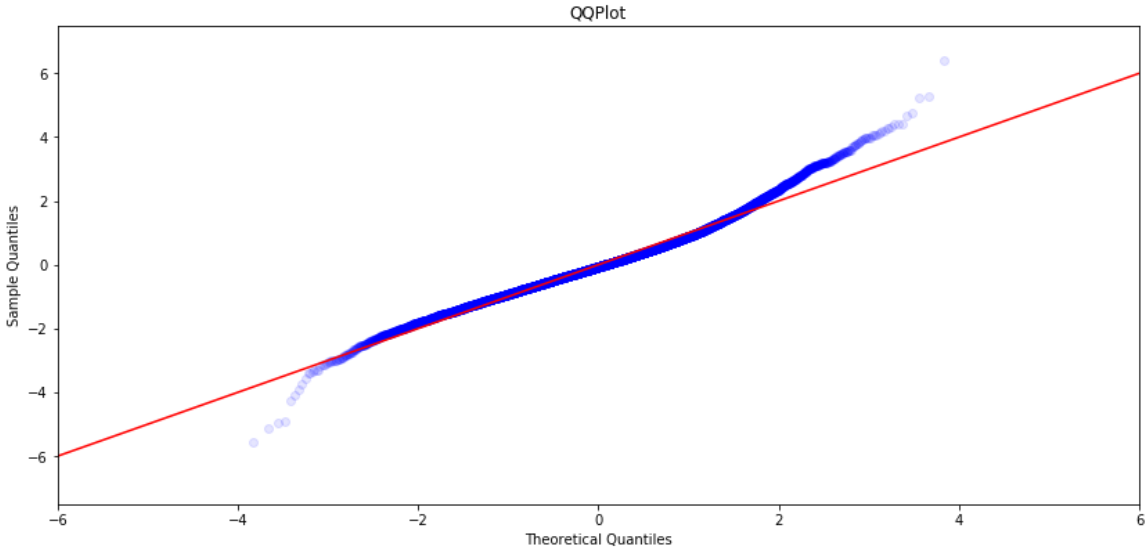
```
2.2e+05
Government       -9458.8903   1123.294     -8.421      0.000   -1.17e+04   -7
257.107
sqft_living15       83.4184      2.035     40.995      0.000      79.430
87.407
Fire             -1876.5335    691.415     -2.714      0.007   -3231.785    -
521.282
bathroomsx4       9467.3613    456.544     20.737      0.000    8572.484
1.04e+04
Cultural          7159.4465    641.956     11.153      0.000    5901.140    8
417.753
Public_Gathering  3386.1507   1535.220      2.206      0.027     376.944    6
395.358
condition         2.491e+04   1319.949     18.870      0.000    2.23e+04
2.75e+04
zipcode           -126.0784     19.886     -6.340      0.000    -165.058
-87.099
date                88.2093      6.897     12.789      0.000      74.690
101.728
==============================================================================
=
Omnibus:                     1008.800   Durbin-Watson:                   2.01
8
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1688.49
2
Skew:                           0.502   Prob(JB):                         0.0
0
Kurtosis:                       4.246   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.
Number of features: 29
sqft_lot15
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.75
0
Model:                            OLS   Adj. R-squared:                  0.74
9
Method:                 Least Squares   F-statistic:                     163
0.
Date:                Tue, 01 Dec 2020   Prob (F-statistic):               0.0
0
Time:                        10:51:56   Log-Likelihood:            -2.0427e+0
5
No. Observations:               15824   AIC:                          4.086e+0
5
Df Residuals:                   15794   BIC:                          4.088e+0
5
```

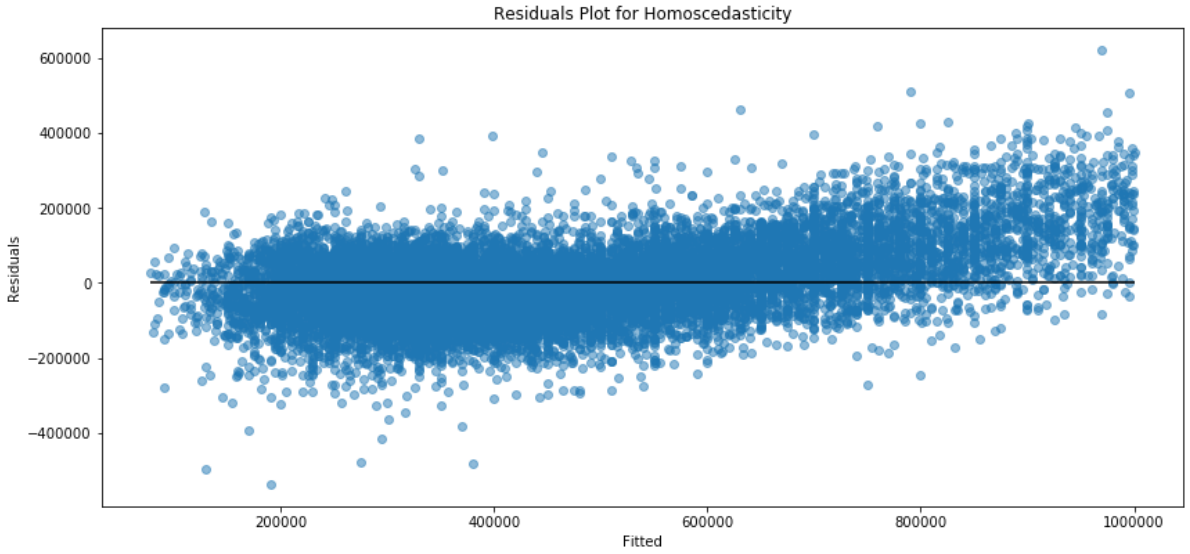
```
Df Model:                           29
Covariance Type:            nonrobust
==============================================================================
=======
                      coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------------------
-------
Intercept        -7.301e+06   2.07e+06     -3.526      0.000   -1.14e+07   -
3.24e+06
sqft_lot             0.2972      0.038      7.757      0.000       0.222
0.372
long              6.567e+04   1.01e+04      6.475      0.000    4.58e+04
8.56e+04
lat               6.319e+05   9004.627     70.170      0.000    6.14e+05
6.5e+05
sqft_basement       13.9968      2.474      5.658      0.000       9.148
18.846
Lodging           6507.6924    656.628      9.911      0.000    5220.627    7
794.758
Campground        3435.5814    268.765     12.783      0.000    2908.771    3
962.392
Access_Point     -6496.8070    288.633    -22.509      0.000   -7062.560   -5
931.054
grade             6.767e+04   1233.379     54.869      0.000    6.53e+04
7.01e+04
Gated_w_Building  8492.6741    241.525     35.163      0.000    8019.258    8
966.091
Airport           2698.6073    302.562      8.919      0.000    2105.552    3
291.663
Commercial_Farm  -3763.7394    459.463     -8.192      0.000   -4664.340   -2
863.139
sqft_lot15        7687.8163   1784.827      4.307      0.000    4189.352
1.12e+04
yr_built         -1224.7106     43.518    -28.142      0.000   -1310.011   -1
139.410
view              2.332e+04   1407.612     16.567      0.000    2.06e+04
2.61e+04
Educational      -1689.9630    795.913     -2.123      0.034   -3250.043    -
129.883
Abandoned        -4849.2504    372.909    -13.004      0.000   -5580.194   -4
118.307
bedrooms          9203.1772   1094.638      8.408      0.000    7057.562
1.13e+04
floorsx2          1.177e+04   1131.138     10.403      0.000    9550.031
1.4e+04
Police            2230.2907    254.554      8.762      0.000    1731.335    2
729.246
waterfront        1.905e+05   1.62e+04     11.738      0.000    1.59e+05
2.22e+05
Government       -8533.5364   1127.541     -7.568      0.000   -1.07e+04   -6
323.427
sqft_living15       86.6055      2.061     42.025      0.000      82.566
90.645
Fire             -1608.8180    694.084     -2.318      0.020   -2969.303    -
248.334
bathroomsx4       9485.9348    458.235     20.701      0.000    8587.741
```

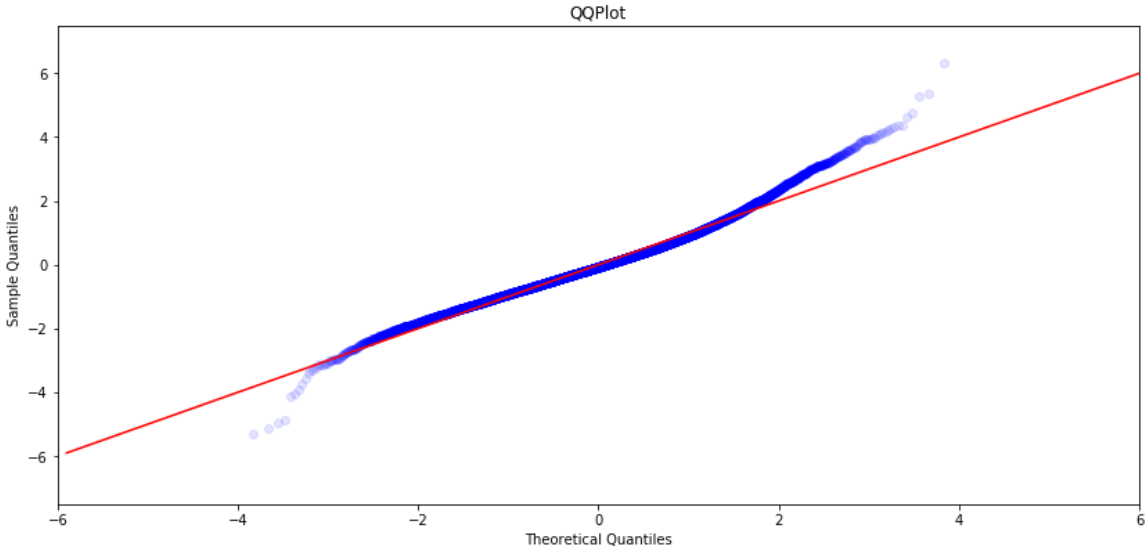
```
1.04e+04
Cultural          6971.2841    643.895     10.827      0.000    5709.177    8
233.391
Public_Gathering  4343.6041   1547.506      2.807      0.005    1310.316    7
376.892
condition         2.486e+04   1326.167     18.743      0.000    2.23e+04
2.75e+04
zipcode           -131.7396     19.947     -6.604      0.000    -170.839
-92.641
date                86.6241      6.922     12.514      0.000      73.056
100.192
==============================================================================
=
Omnibus:                     1044.153   Durbin-Watson:                   2.01
6
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1785.76
3
Skew:                           0.509   Prob(JB):                         0.0
0
Kurtosis:                       4.293   Cond. No.                     2.63e+0
8
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
[2] The condition number is large, 2.63e+08. This might indicate that there a
re
strong multicollinearity or other numerical problems.
```

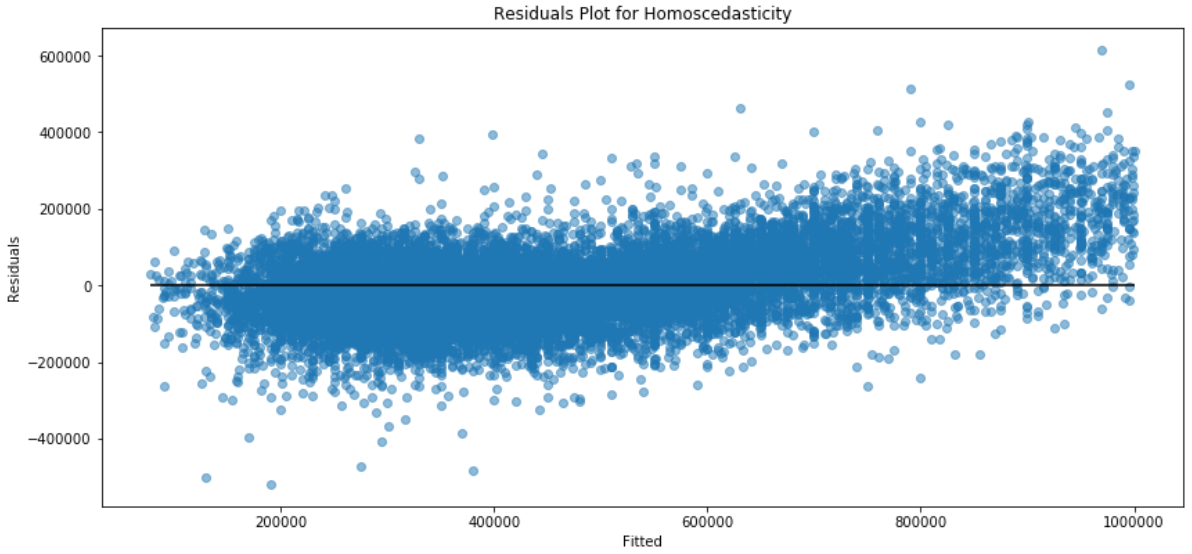

QQPlot
Sample Quantiles
Theoretical Quantiles
6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6

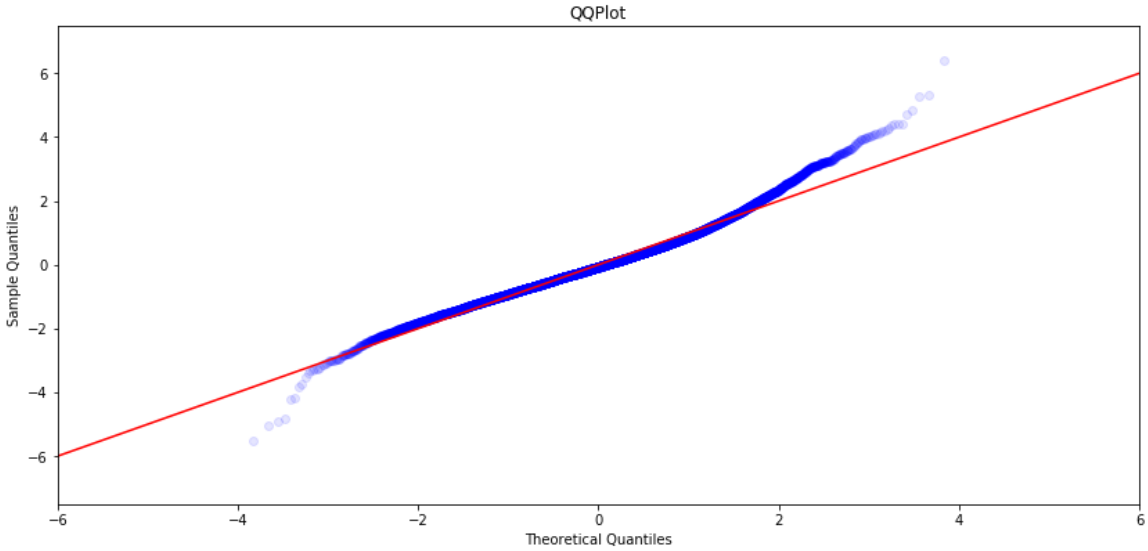

Residuals Plot for Homoscedasticity
Residuals
Fitted
600000
400000
200000
0
−200000
−400000
−600000
200000
400000
600000
800000
1000000

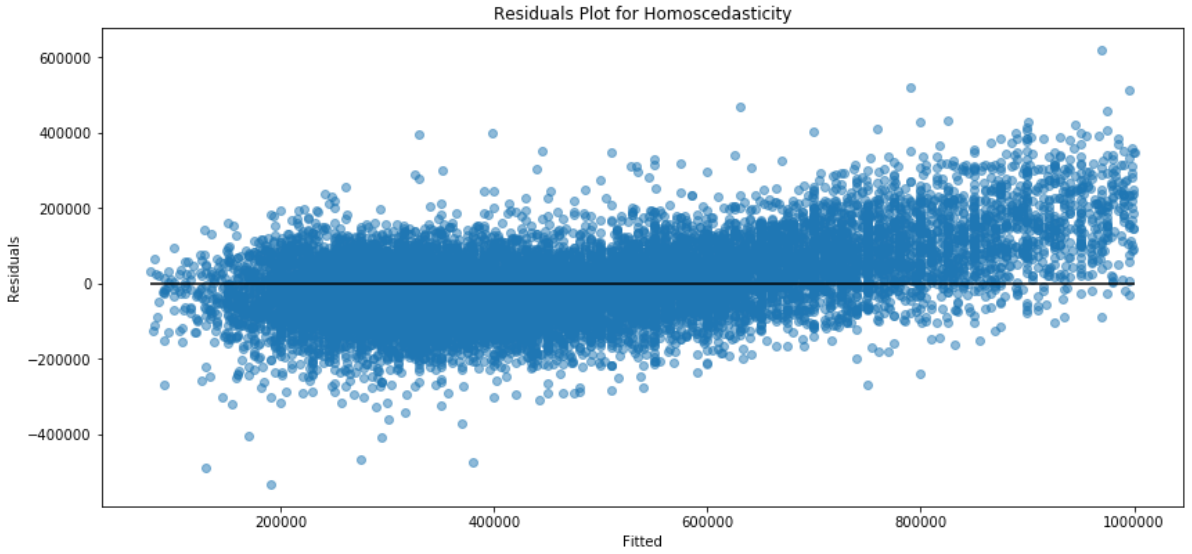

QQPlot
Sample Quantiles
Theoretical Quantiles
6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6

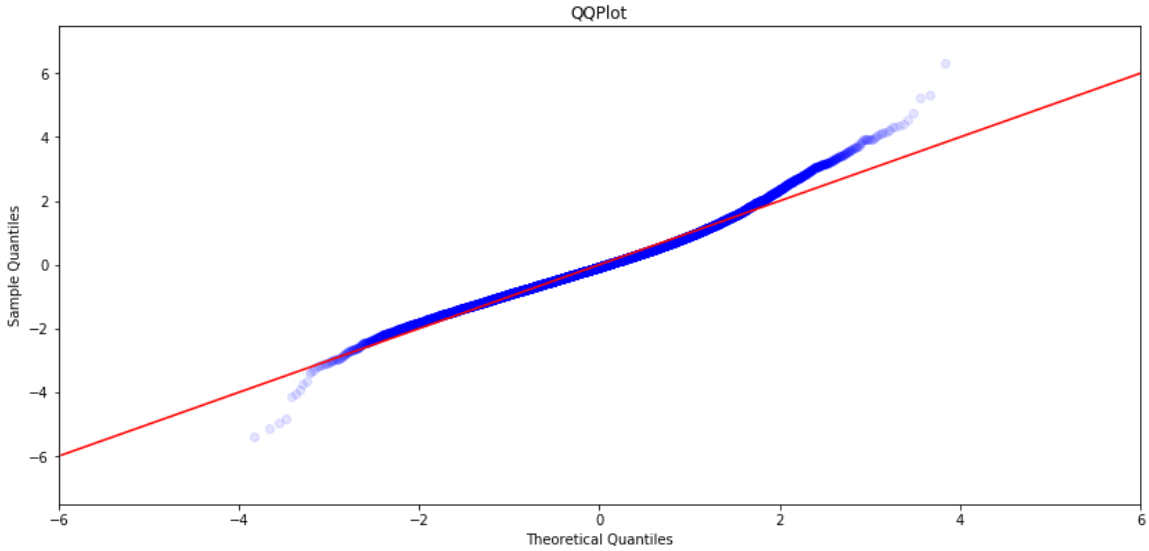

Residuals Plot for Homoscedasticity
Residuals
Fitted
600000
400000
200000
0
−200000
−400000
200000
400000
600000
800000
1000000

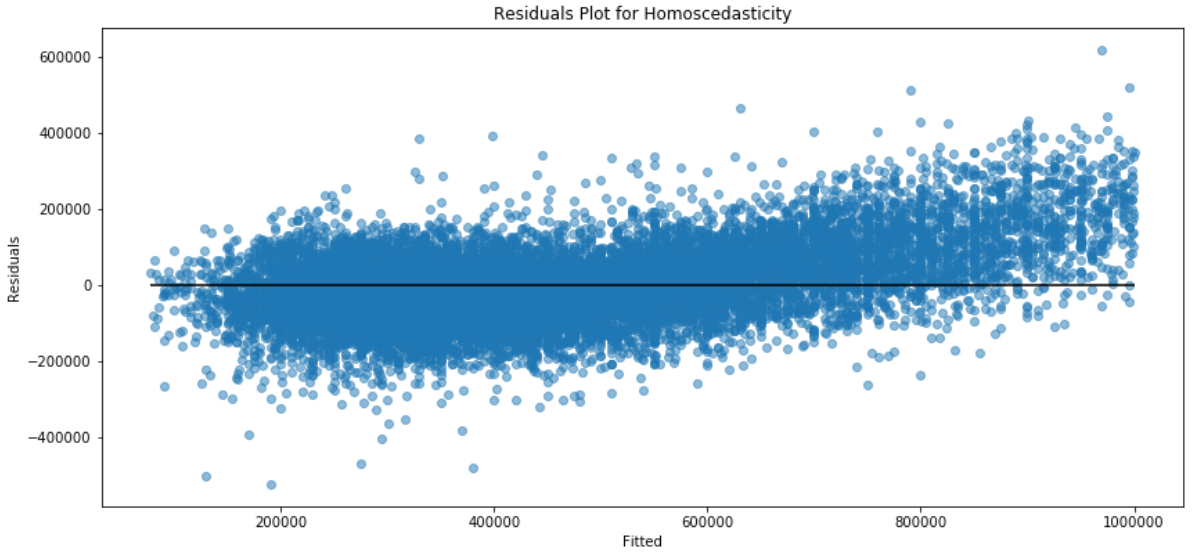

QQPlot
Sample Quantiles
Theoretical Quantiles
6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6

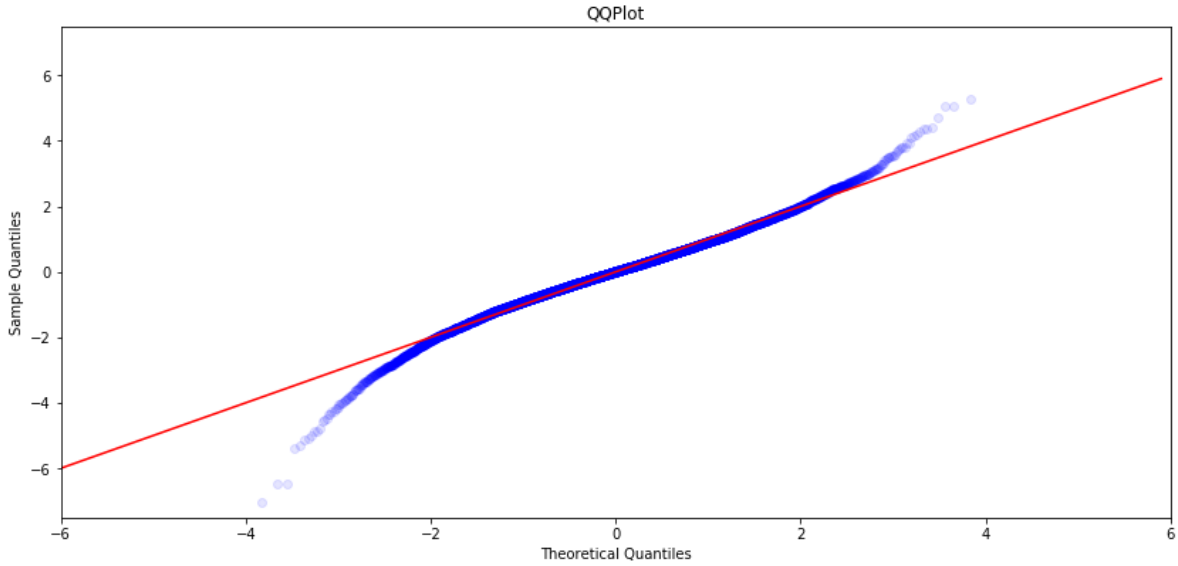

Residuals Plot for Homoscedasticity
Residuals
Fitted
600000
400000
200000
0
−200000
−400000
200000
400000
600000
800000
1000000

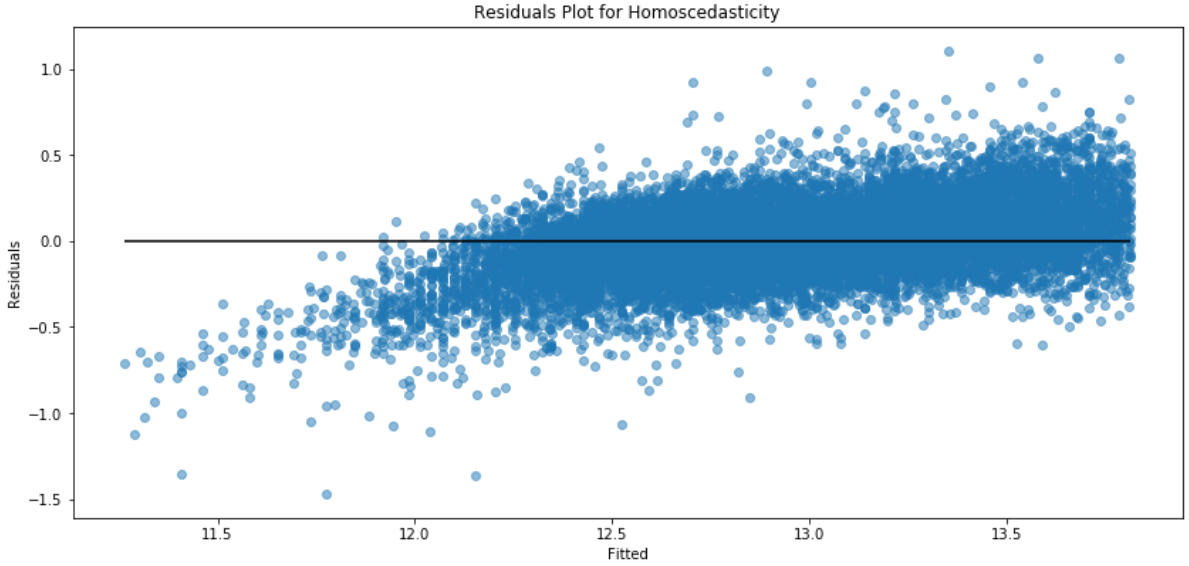

QQPlot
Sample Quantiles
Theoretical Quantiles
6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6

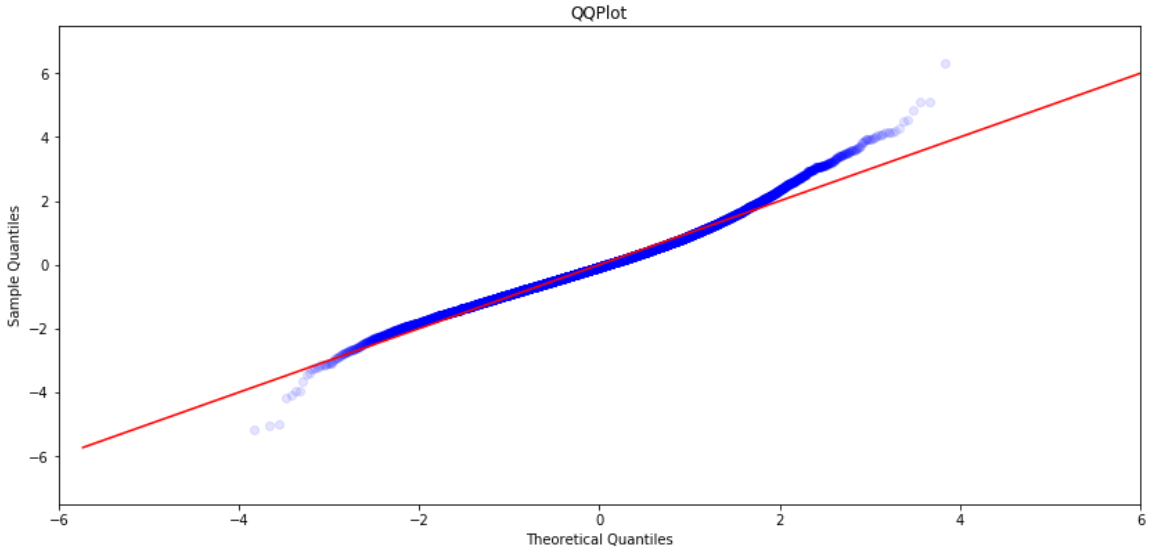

Residuals Plot for Homoscedasticity
Residuals
Fitted
600000
400000
200000
0
−200000
−400000
200000
400000
600000
800000
1000000

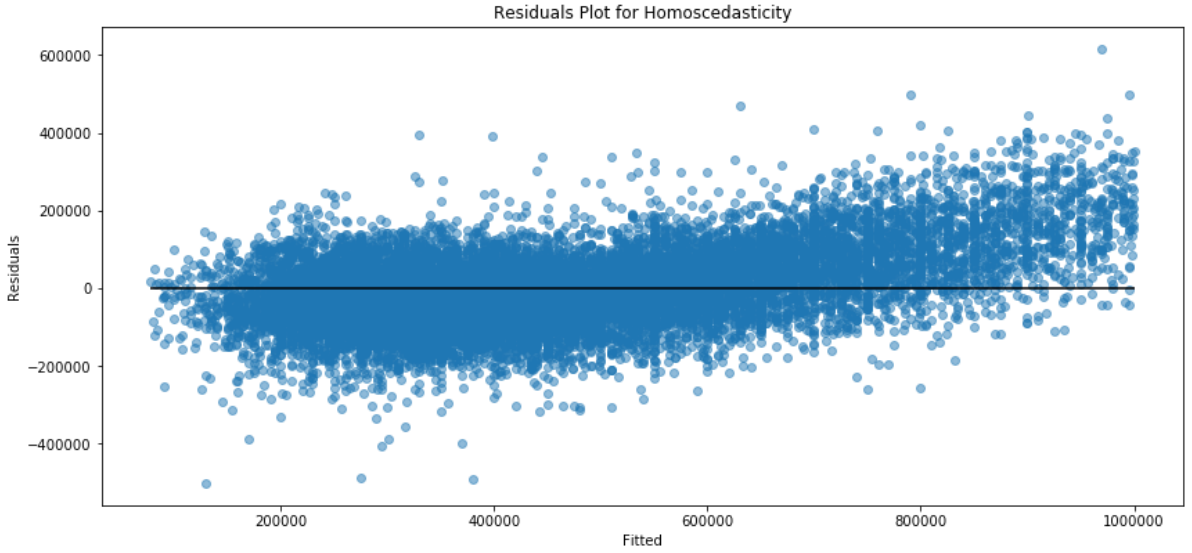

QQPlot
Sample Quantiles
Theoretical Quantiles
6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6

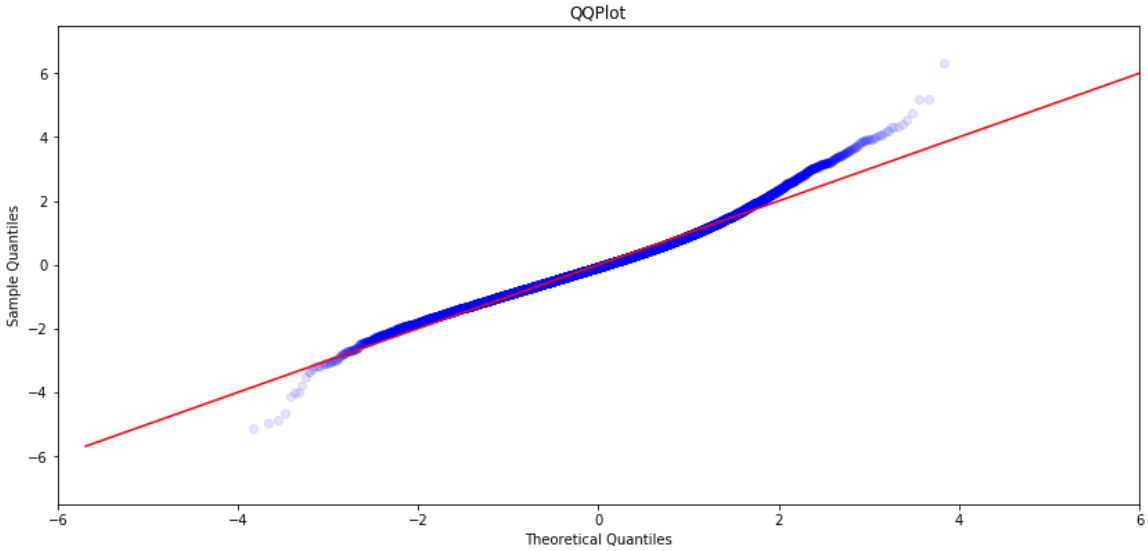

Residuals Plot for Homoscedasticity
Residuals
Fitted
600000
400000
200000
0
−200000
−400000
200000
400000
600000
800000
1000000

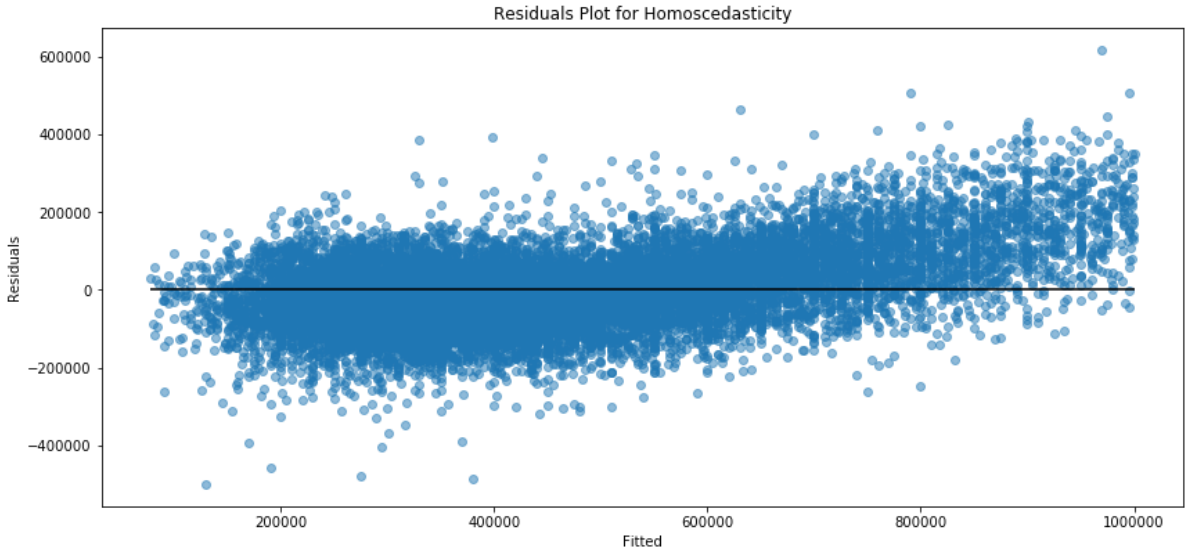

QQPlot
Sample Quantiles
Theoretical Quantiles
6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6


Residuals Plot for Homoscedasticity
Residuals
Fitted
1.0
0.5
0.0
−0.5
−1.0
−1.5
11.5
12.0
12.5
13.0
13.5


QQPlot
Sample Quantiles
Theoretical Quantiles
6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6


Residuals Plot for Homoscedasticity
Residuals
Fitted
600000
400000
200000
0
−200000
−400000
200000
400000
600000
800000
1000000





Scaled-Log

Again, as above, but transforming with my minmax_plus function first

In [24]:
```
data_t = data_train.copy()
multicollinearity_threshold=0.7
alpha=0.1
cand_log = data_t.drop(['lat', 'long', 'waterfront', 'zipcode'], axis=1).colum
ns

for feat in cand_log:
    data_t = data_train.copy()
    to_log = [feat]
    data_t = minmax_plus(data_t, to_log)
    data_t = log(data_t, to_log)
    x_cols = data_t.drop([outcome], axis=1).columns
    x_cols = multicoll_remove(data_t, x_cols, multicollinearity_threshold)
    x_cols = simple_selector(data_t, x_cols)


    results = model(data_t, x_cols)

    if results.rsquared_adj > 0.749:
        metrics(data_t, results, x_cols)
        print(feat)
        print(results.summary())
```

Number of features: 26
Gate_wo_Building

OLS Regression Results

Dep. Variable:	price	R-squared:	0.753
Model:	OLS	Adj. R-squared:	0.753
Method:	Least Squares	F-statistic:	1853.
Date:	Tue, 01 Dec 2020	Prob (F-statistic):	0.00
Time:	10:52:21	Log-Likelihood:	-2.0416e+05
No. Observations:	15824	AIC:	4.084e+05
Df Residuals:	15797	BIC:	4.086e+05
Df Model:	26		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-2.746e+07	4.55e+05	-60.376	0.000	-2.84e+07	-2.66e+07
sqft_lot	0.3673	0.035	10.476	0.000	0.299	0.436
lat	6.2e+05	9195.710	67.426	0.000	6.02e+05	6.38e+05
sqft_basement	14.9321	2.450	6.095	0.000	10.130	19.735
Lodging	7348.0420	654.644	11.224	0.000	6064.865	8631.219
Campground	2994.9030	250.107	11.974	0.000	2504.665	3485.141
Access_Point	-5624.6184	279.829	-20.100	0.000	-6173.115	-5076.122
grade	6.749e+04	1218.612	55.381	0.000	6.51e+04	6.99e+04
Gated_w_Building	9412.0339	254.690	36.955	0.000	8912.813	9911.254
Airport	1380.0970	307.297	4.491	0.000	777.759	1982.435
Commercial_Farm	-3324.4850	427.511	-7.776	0.000	-4162.455	-2486.515
yr_built	-1270.1994	42.158	-30.129	0.000	-1352.835	-1187.564
Cemetery	-3933.8107	391.296	-10.053	0.000	-4700.795	-3166.826
view	2.371e+04	1391.685	17.034	0.000	2.1e+04	2.64e+04
Abandoned	-3857.7460	356.775	-10.813	0.000	-4557.065	-3158.427

```
bedrooms          1.002e+04   1079.341      9.281      0.000    7902.246
1.21e+04
floorsx2          1.018e+04   1068.396      9.524      0.000    8081.599
1.23e+04
Police            3744.5786    271.293     13.803      0.000    3212.814    4
276.343
waterfront        1.919e+05   1.61e+04     11.940      0.000     1.6e+05
2.23e+05
Government       -8213.8497   1109.208     -7.405      0.000   -1.04e+04   -6
039.676
sqft_living15       86.4772      1.988     43.491      0.000      82.580
90.375
bathroomsx4       9409.6581    454.597     20.699      0.000    8518.595
1.03e+04
Cultural          5706.1449    643.786      8.863      0.000    4444.252    6
968.038
Seasonal_Home    -2873.2085    149.809    -19.179      0.000   -3166.852   -2
579.565
Public_Gathering  6343.8583   1459.112      4.348      0.000    3483.833    9
203.884
condition          2.46e+04   1310.796     18.764      0.000     2.2e+04
2.72e+04
date                 85.0915      6.869     12.388      0.000      71.627
98.556
==============================================================================
=
Omnibus:                     1142.766   Durbin-Watson:                   2.01
3
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2116.95
1
Skew:                           0.522   Prob(JB):                         0.0
0
Kurtosis:                       4.456   Cond. No.                     1.54e+0
7
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
[2] The condition number is large, 1.54e+07. This might indicate that there a
re
strong multicollinearity or other numerical problems.
Number of features: 27
Police
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.75
4
Model:                            OLS   Adj. R-squared:                  0.75
4
Method:                 Least Squares   F-statistic:                     179
7.
Date:                Tue, 01 Dec 2020   Prob (F-statistic):               0.0
0
Time:                        10:52:22   Log-Likelihood:           -2.0412e+0
```

```
5
No. Observations:               15824   AIC:                         4.083e+0
5
Df Residuals:                   15796   BIC:                         4.085e+0
5
Df Model:                          27
Covariance Type:            nonrobust
================================================================================
=======
                       coef    std err          t      P>|t|      [0.025
0.975]
--------------------------------------------------------------------------------
-------
Intercept        -2.294e+07   5.55e+05    -41.346      0.000    -2.4e+07   -
2.19e+07
sqft_lot             0.3581      0.035     10.236      0.000       0.290
0.427
lat                5.239e+05   1.13e+04     46.445      0.000    5.02e+05
5.46e+05
sqft_basement        13.3165      2.438      5.461      0.000       8.537
18.096
Lodging            7381.3131    647.762     11.395      0.000    6111.626    8
651.000
Campground         3151.4442    257.027     12.261      0.000    2647.642    3
655.247
Access_Point      -5919.1578    273.840    -21.615      0.000   -6455.916   -5
382.400
grade              6.771e+04   1216.316     55.666      0.000    6.53e+04
7.01e+04
Gated_w_Building    1.04e+04    262.077     39.677      0.000    9884.605
1.09e+04
Airport             998.1649    311.024      3.209      0.001     388.523    1
607.807
Commercial_Farm   -3367.8959    447.910     -7.519      0.000   -4245.852   -2
489.940
yr_built          -1257.2336     41.978    -29.950      0.000   -1339.515   -1
174.952
Cemetery          -5891.6384    438.641    -13.432      0.000   -6751.426   -5
031.851
view               2.337e+04   1386.890     16.849      0.000    2.06e+04
2.61e+04
Educational        3635.9967    789.276      4.607      0.000    2088.925    5
183.068
Abandoned         -3100.5062    352.279     -8.801      0.000   -3791.014   -2
409.998
bedrooms           9903.5590   1076.157      9.203      0.000    7794.169
1.2e+04
floorsx2           1.005e+04   1066.448      9.422      0.000    7957.523
1.21e+04
Police             2.382e+04   1592.341     14.960      0.000    2.07e+04
2.69e+04
waterfront         1.899e+05    1.6e+04     11.853      0.000    1.58e+05
2.21e+05
Government        -7957.3429   1105.545     -7.198      0.000   -1.01e+04   -5
790.348
Gate_wo_Building -3452.1522    171.447    -20.135      0.000   -3788.208   -3
116.096
```

```
sqft_living15       86.5795      1.982     43.687      0.000      82.695
90.464
bathroomsx4       9436.8666    453.377     20.815      0.000    8548.196
1.03e+04
Cultural          6167.5707    637.862      9.669      0.000    4917.289    7
417.852
Public_Gathering  3373.5272   1511.231      2.232      0.026     411.342    6
335.712
condition         2.432e+04   1307.404     18.600      0.000    2.18e+04
2.69e+04
date                85.1927      6.851     12.435      0.000      71.764
98.621
==============================================================================
=
Omnibus:                     1119.938   Durbin-Watson:                   2.01
4
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2091.11
4
Skew:                           0.511   Prob(JB):                         0.0
0
Kurtosis:                       4.458   Cond. No.                     1.88e+0
7
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
[2] The condition number is large, 1.88e+07. This might indicate that there a
re
strong multicollinearity or other numerical problems.
Number of features: 28
Public_Gathering
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.75
0
Model:                            OLS   Adj. R-squared:                  0.74
9
Method:                 Least Squares   F-statistic:                     168
8.
Date:                Tue, 01 Dec 2020   Prob (F-statistic):               0.0
0
Time:                        10:52:23   Log-Likelihood:            -2.0428e+0
5
No. Observations:               15824   AIC:                         4.086e+0
5
Df Residuals:                   15795   BIC:                         4.088e+0
5
Df Model:                          28
Covariance Type:            nonrobust
==============================================================================
=======
                       coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------------------
```

```
-------
Intercept        -7.305e+06   2.07e+06     -3.528      0.000   -1.14e+07   -
3.25e+06
sqft_lot             0.3573      0.035     10.131      0.000       0.288
0.426
long              6.547e+04   1.01e+04      6.465      0.000    4.56e+04
8.53e+04
lat               6.315e+05   8999.773     70.166      0.000    6.14e+05
6.49e+05
sqft_basement       13.3538      2.468      5.411      0.000       8.516
18.191
Lodging           6541.8299    653.131     10.016      0.000    5261.619    7
822.040
Campground        3168.9617    261.223     12.131      0.000    2656.935    3
680.988
Access_Point     -6496.4408    288.536    -22.515      0.000   -7062.004   -5
930.877
grade             6.797e+04   1229.693     55.270      0.000    6.56e+04
7.04e+04
Gated_w_Building  8570.5430    243.191     35.242      0.000    8093.860    9
047.226
Airport           2648.9983    302.636      8.753      0.000    2055.797    3
242.200
Commercial_Farm  -3815.1233    458.895     -8.314      0.000   -4714.609   -2
915.637
yr_built         -1265.4080     42.657    -29.664      0.000   -1349.022   -1
181.794
view              2.327e+04   1407.730     16.527      0.000    2.05e+04
2.6e+04
Educational      -1421.5144    775.779     -1.832      0.067   -2942.131
99.102
Abandoned        -4523.9182    364.039    -12.427      0.000   -5237.476   -3
810.360
bedrooms          9710.1386   1087.743      8.927      0.000    7578.038
1.18e+04
floorsx2          1.046e+04   1079.279      9.687      0.000    8339.752
1.26e+04
Police            2170.7448    254.170      8.541      0.000    1672.543    2
668.947
waterfront        1.927e+05   1.62e+04     11.883      0.000    1.61e+05
2.24e+05
Government       -7935.8019   1091.587     -7.270      0.000   -1.01e+04   -5
796.168
sqft_living15       88.3469      2.004     44.080      0.000      84.418
92.275
Fire             -1317.6027    688.355     -1.914      0.056   -2666.857
31.651
bathroomsx4       9464.3740    458.157     20.657      0.000    8566.334
1.04e+04
Cultural          6958.9312    642.775     10.826      0.000    5699.018    8
218.844
Public_Gathering  9729.9122   1855.266      5.244      0.000    6093.379
1.34e+04
condition         2.502e+04   1325.044     18.883      0.000    2.24e+04
2.76e+04
zipcode           -130.4007     19.949     -6.537      0.000    -169.504
-91.298
```

```
date                85.8571      6.919     12.408      0.000      72.295
99.420
==============================================================================
=
Omnibus:                     1051.241   Durbin-Watson:                   2.01
5
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1832.50
5
Skew:                           0.506   Prob(JB):                         0.0
0
Kurtosis:                       4.325   Cond. No.                     2.63e+0
8
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
[2] The condition number is large, 2.63e+08. This might indicate that there a
re
strong multicollinearity or other numerical problems.
Number of features: 28
Seasonal_Home
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.75
3
Model:                            OLS   Adj. R-squared:                  0.75
3
Method:                 Least Squares   F-statistic:                     171
9.
Date:                Tue, 01 Dec 2020   Prob (F-statistic):               0.0
0
Time:                        10:52:24   Log-Likelihood:            -2.0417e+0
5
No. Observations:               15824   AIC:                             4.084e+0
5
Df Residuals:                   15795   BIC:                             4.086e+0
5
Df Model:                          28
Covariance Type:            nonrobust
==============================================================================
=======
                       coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------------------
-------
Intercept        -1.947e+07   1.94e+06    -10.009      0.000   -2.33e+07   -
1.57e+07
sqft_lot             0.3601      0.035     10.263      0.000       0.291
0.429
lat               5.298e+05   1.19e+04     44.414      0.000    5.06e+05
5.53e+05
sqft_basement       13.2714      2.448      5.422      0.000       8.473
18.069
Lodging           7285.7516    655.212     11.120      0.000    6001.462    8
```

```
570.042
Campground        3409.8446    260.179     13.106      0.000    2899.864    3
919.825
Access_Point     -6275.6979    280.531    -22.371      0.000   -6825.571   -5
725.824
grade             6.792e+04   1219.792     55.685      0.000    6.55e+04
7.03e+04
Gated_w_Building  9812.3538    263.705     37.210      0.000    9295.462
1.03e+04
Airport           1580.7705    308.529      5.124      0.000     976.019    2
185.522
Commercial_Farm  -3241.2191    450.394     -7.196      0.000   -4124.043   -2
358.395
yr_built         -1256.7328     42.343    -29.680      0.000   -1339.729   -1
173.736
Cemetery         -5557.2367    462.638    -12.012      0.000   -6464.060   -4
650.414
view              2.365e+04   1397.392     16.927      0.000    2.09e+04
2.64e+04
Educational       3361.5359    799.653      4.204      0.000    1794.124    4
928.948
Abandoned        -3414.5485    372.512     -9.166      0.000   -4144.715   -2
684.382
bedrooms          9934.7421   1080.365      9.196      0.000    7817.103
1.21e+04
floorsx2          9856.1471   1071.554      9.198      0.000    7755.780
1.2e+04
Police            3014.3116    267.640     11.263      0.000    2489.707    3
538.917
waterfront        1.908e+05   1.61e+04     11.861      0.000    1.59e+05
2.22e+05
Government       -7482.6255   1110.452     -6.738      0.000   -9659.237   -5
306.014
Gate_wo_Building -3054.3307    185.220    -16.490      0.000   -3417.383   -2
691.278
sqft_living15       87.3176      1.988     43.929      0.000      83.422
91.214
bathroomsx4       9419.3499    454.933     20.705      0.000    8527.630
1.03e+04
Cultural          6369.4055    639.744      9.956      0.000    5115.435    7
623.376
Public_Gathering  4430.6843   1512.406      2.930      0.003    1466.197    7
395.172
condition         2.471e+04   1316.117     18.772      0.000    2.21e+04
2.73e+04
zipcode            -37.9895     20.569     -1.847      0.065     -78.307
2.328
date                85.7596      6.872     12.480      0.000      72.290
99.229
==============================================================================
=
Omnibus:                     1137.198   Durbin-Watson:                   2.01
3
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2104.82
4
Skew:                           0.520   Prob(JB):                         0.0
0
```

```
Kurtosis:                       4.453   Cond. No.                     2.49e+0
8
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
[2] The condition number is large, 2.49e+08. This might indicate that there a
re
strong multicollinearity or other numerical problems.
Number of features: 28
floorsx2
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.75
0
Model:                            OLS   Adj. R-squared:                  0.75
0
Method:                 Least Squares   F-statistic:                     169
6.
Date:                Tue, 01 Dec 2020   Prob (F-statistic):               0.0
0
Time:                        10:52:26   Log-Likelihood:            -2.0425e+0
5
No. Observations:               15824   AIC:                         4.086e+0
5
Df Residuals:                   15795   BIC:                         4.088e+0
5
Df Model:                          28
Covariance Type:            nonrobust
==============================================================================
=======
                       coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------------------
-------
Intercept        -7.573e+06   2.07e+06     -3.666      0.000   -1.16e+07   -
3.52e+06
sqft_lot             0.3682      0.035     10.434      0.000       0.299
0.437
long              6.009e+04   1.01e+04      5.927      0.000    4.02e+04
8e+04
lat               6.311e+05   8979.285     70.282      0.000    6.13e+05
6.49e+05
sqft_basement       17.2950      2.486      6.957      0.000      12.422
22.168
Lodging           6558.6192    655.080     10.012      0.000    5274.587    7
842.651
Campground        3110.6681    259.759     11.975      0.000    2601.511    3
619.825
Access_Point     -6597.8128    288.419    -22.876      0.000   -7163.146   -6
032.479
grade              6.76e+04   1227.419     55.073      0.000    6.52e+04
7e+04
Gated_w_Building  8420.2118    240.994     34.939      0.000    7947.835    8
```

```
892.588
Airport           2778.5137    302.185      9.195      0.000    2186.197    3
370.831
Commercial_Farm  -3848.4598    458.483     -8.394      0.000   -4747.139   -2
949.781
yr_built         -1210.2459     41.259    -29.333      0.000   -1291.117   -1
129.374
view              2.369e+04   1405.243     16.858      0.000    2.09e+04
2.64e+04
Educational      -1696.0451    794.304     -2.135      0.033   -3252.972    -
139.118
Abandoned        -4384.0584    363.515    -12.060      0.000   -5096.589   -3
671.528
bedrooms          8843.9166   1090.209      8.112      0.000    6706.982
1.1e+04
floorsx2          7607.2307    601.024     12.657      0.000    6429.155    8
785.306
Police            2161.2956    253.995      8.509      0.000    1663.437    2
659.155
waterfront        1.923e+05   1.62e+04     11.878      0.000    1.61e+05
2.24e+05
Government       -7998.6207   1116.330     -7.165      0.000   -1.02e+04   -5
810.487
sqft_living15       87.6780      2.003     43.764      0.000      83.751
91.605
Fire             -1424.9217    691.588     -2.060      0.039   -2780.512
-69.331
bathroomsx4       8934.0172    458.380     19.490      0.000    8035.539    9
832.495
Cultural          6884.7108    642.606     10.714      0.000    5625.131    8
144.291
Public_Gathering  5608.5868   1527.888      3.671      0.000    2613.751    8
603.422
condition         2.549e+04   1322.982     19.266      0.000    2.29e+04
2.81e+04
zipcode           -134.9056     19.917     -6.774      0.000    -173.945
-95.867
date                86.1901      6.908     12.476      0.000      72.649
99.731
==============================================================================
=
Omnibus:                     1069.496   Durbin-Watson:                   2.01
6
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1883.37
7
Skew:                           0.510   Prob(JB):                         0.0
0
Kurtosis:                       4.348   Cond. No.                     2.63e+0
8
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
[2] The condition number is large, 2.63e+08. This might indicate that there a
re
```

strong multicollinearity or other numerical problems.
Number of features: 25
sqft_above

OLS Regression Results

Dep. Variable:	price	R-squared:	0.763
Model:	OLS	Adj. R-squared:	0.763
Method:	Least Squares	F-statistic:	2035.
Date:	Tue, 01 Dec 2020	Prob (F-statistic):	0.00
Time:	10:52:28	Log-Likelihood:	-2.0384e+05
No. Observations:	15824	AIC:	4.077e+05
Df Residuals:	15798	BIC:	4.079e+05
Df Model:	25		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-4.393e+06	2.01e+06	-2.191	0.028	-8.32e+06	-4.63e+05
sqft_lot	0.2400	0.034	6.970	0.000	0.172	0.307
long	5.401e+04	9817.071	5.502	0.000	3.48e+04	7.33e+04
lat	6.784e+05	8686.287	78.098	0.000	6.61e+05	6.95e+05
sqft_basement	-38.7410	2.303	-16.823	0.000	-43.255	-34.227
Lodging	6702.6424	627.148	10.687	0.000	5473.360	7931.924
Campground	3496.0298	245.753	14.226	0.000	3014.327	3977.733
Access_Point	-7257.0894	274.789	-26.410	0.000	-7795.707	-6718.472
grade	5.86e+04	1234.149	47.484	0.000	5.62e+04	6.1e+04
Gated_w_Building	9300.7427	233.521	39.828	0.000	8843.016	9758.470
Airport	2564.2873	293.534	8.736	0.000	1988.928	3139.647
Commercial_Farm	-4247.3410	438.270	-9.691	0.000	-5106.399	-3388.283
yr_built	-836.1566	36.429	-22.953	0.000	-907.561	-764.752
view	2.906e+04	1355.124	21.445	0.000	2.64e+04	3.17e+04
Educational	-1913.1782	772.877	-2.475	0.013	-3428.106	-

```
398.250
Abandoned        -5210.5328    348.492    -14.952      0.000   -5893.616   -4
527.450
bedrooms         -6167.1312   1113.700     -5.538      0.000   -8350.111   -3
984.151
sqft_living        123.8887      1.933     64.093      0.000     120.100
127.678
Police            2343.7199    246.898      9.493      0.000    1859.772    2
827.668
waterfront        1.841e+05   1.58e+04     11.691      0.000    1.53e+05
2.15e+05
Government       -7746.1247   1083.793     -7.147      0.000   -9870.484   -5
621.766
Cultural          9659.0892    622.972     15.505      0.000    8437.993
1.09e+04
Public_Gathering  4194.3984   1475.719      2.842      0.004    1301.820    7
086.977
condition         2.473e+04   1283.207     19.276      0.000    2.22e+04
2.72e+04
zipcode           -203.9571     19.351    -10.540      0.000    -241.887    -
166.027
date                82.5722      6.727     12.274      0.000      69.386
95.759
==============================================================================
=
Omnibus:                     1030.279   Durbin-Watson:                   2.02
8
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1980.48
4
Skew:                           0.468   Prob(JB):                         0.0
0
Kurtosis:                       4.458   Cond. No.                     2.62e+0
8
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
[2] The condition number is large, 2.62e+08. This might indicate that there a
re
strong multicollinearity or other numerical problems.
Number of features: 28
sqft_living
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.76
4
Model:                            OLS   Adj. R-squared:                  0.76
4
Method:                 Least Squares   F-statistic:                     182
6.
Date:                Tue, 01 Dec 2020   Prob (F-statistic):               0.0
0
Time:                        10:52:29   Log-Likelihood:           -2.0381e+0
5
```

```
No. Observations:               15824   AIC:                          4.077e+0
5
Df Residuals:                   15795   BIC:                          4.079e+0
5
Df Model:                          28
Covariance Type:            nonrobust
==============================================================================
=======
                      coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------------------
-------
Intercept        -5.459e+06   2.01e+06     -2.720      0.007   -9.39e+06   -
1.52e+06
sqft_lot             0.2451      0.034      7.120      0.000       0.178
0.313
long              4.927e+04   9856.232      4.999      0.000       3e+04
6.86e+04
lat               6.787e+05   8672.881     78.258      0.000    6.62e+05
6.96e+05
sqft_basement       72.7064      2.589     28.087      0.000      67.632
77.780
Lodging           6912.2166    636.339     10.862      0.000    5664.919    8
159.514
Campground        3328.1250    253.060     13.152      0.000    2832.099    3
824.151
Access_Point     -7403.6666    280.280    -26.415      0.000   -7953.048   -6
854.286
grade             5.821e+04   1232.958     47.211      0.000    5.58e+04
6.06e+04
sqft_above         117.0263      2.116     55.296      0.000     112.878
121.175
Gated_w_Building  9253.7900    234.241     39.505      0.000    8794.651    9
712.929
Airport           2591.1671    293.361      8.833      0.000    2016.146    3
166.188
Commercial_Farm  -3995.9595    445.846     -8.963      0.000   -4869.869   -3
122.050
yr_built          -981.0001     42.104    -23.299      0.000   -1063.529    -
898.471
view              2.913e+04   1353.064     21.532      0.000    2.65e+04
3.18e+04
Educational      -1815.5921    772.096     -2.352      0.019   -3328.989    -
302.195
Abandoned        -5044.6020    354.215    -14.242      0.000   -5738.903   -4
350.301
bedrooms         -7330.4868   1127.505     -6.502      0.000   -9540.526   -5
120.448
floorsx2         -2262.1440   1073.204     -2.108      0.035   -4365.746    -
158.542
Police            2358.5290    246.799      9.556      0.000    1874.774    2
842.284
waterfront        1.823e+05   1.57e+04     11.586      0.000    1.51e+05
2.13e+05
Government       -7631.7802   1084.357     -7.038      0.000   -9757.244   -5
506.316
Fire             -1084.8496    671.818     -1.615      0.106   -2401.689
```

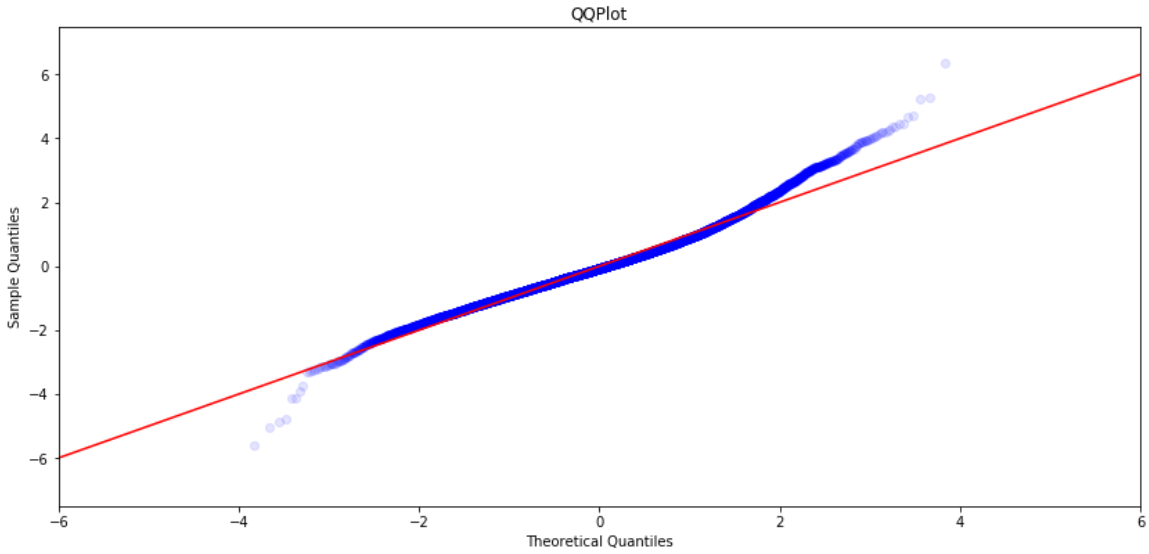
```
231.990
bathroomsx4       4538.3119    460.884      9.847      0.000    3634.927    5
441.697
Cultural          9540.8986    623.645     15.299      0.000    8318.483
1.08e+04
Public_Gathering  4965.9123   1485.187      3.344      0.001    2054.776    7
877.049
condition         2.445e+04   1284.574     19.036      0.000    2.19e+04
2.7e+04
zipcode           -196.2938     19.336    -10.152      0.000    -234.195    -
158.392
date                83.7365      6.717     12.467      0.000      70.571
96.902
==============================================================================
=
Omnibus:                     1032.131   Durbin-Watson:                   2.02
5
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1996.39
8
Skew:                           0.467   Prob(JB):                         0.0
0
Kurtosis:                       4.468   Cond. No.                     2.63e+0
8
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
[2] The condition number is large, 2.63e+08. This might indicate that there a
re
strong multicollinearity or other numerical problems.
Number of features: 28
sqft_lot
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.75
0
Model:                            OLS   Adj. R-squared:                  0.74
9
Method:                 Least Squares   F-statistic:                     169
1.
Date:                Tue, 01 Dec 2020   Prob (F-statistic):               0.0
0
Time:                        10:52:30   Log-Likelihood:            -2.0426e+0
5
No. Observations:               15824   AIC:                         4.086e+0
5
Df Residuals:                   15795   BIC:                         4.088e+0
5
Df Model:                          28
Covariance Type:            nonrobust
==============================================================================
=======
                      coef    std err          t      P>|t|      [0.025
0.975]
```

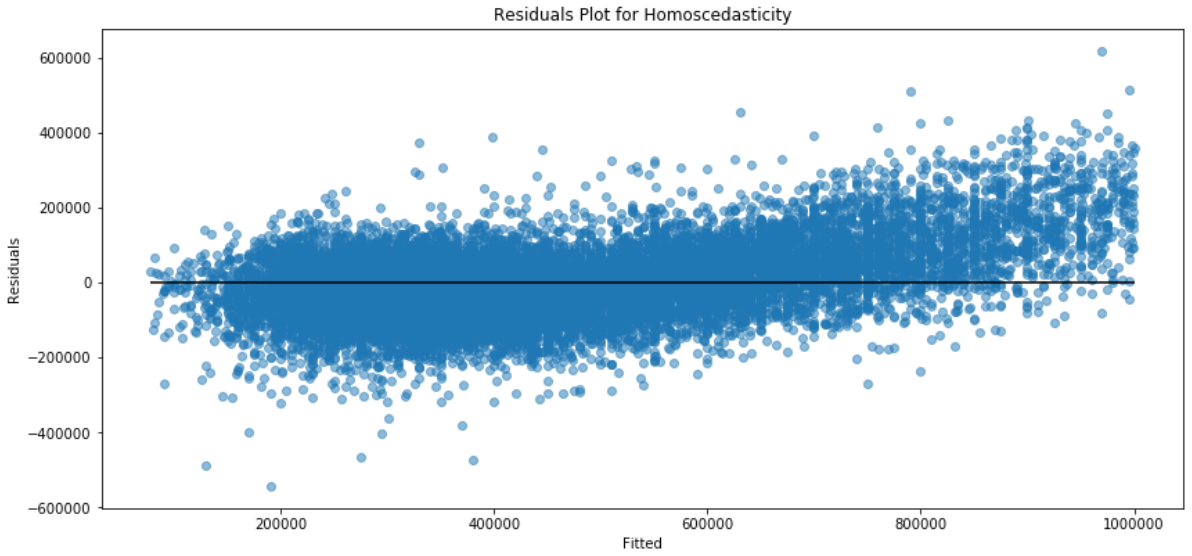
```
------------------------------------------------------------------------------
-------
Intercept        -7.386e+06   2.07e+06     -3.570      0.000   -1.14e+07   -
3.33e+06
sqft_lot          2.995e+04   2512.534     11.921      0.000     2.5e+04
3.49e+04
long              6.677e+04   1.01e+04      6.599      0.000    4.69e+04
8.66e+04
lat                6.32e+05   8994.422     70.265      0.000    6.14e+05
6.5e+05
sqft_basement       13.3946      2.467      5.430      0.000       8.559
18.230
Lodging           6521.8394    655.963      9.942      0.000    5236.077    7
807.602
Campground        3524.3274    263.330     13.384      0.000    3008.171    4
040.484
Access_Point     -6544.5677    288.629    -22.675      0.000   -7110.313   -5
978.822
grade             6.732e+04   1232.307     54.628      0.000    6.49e+04
6.97e+04
Gated_w_Building  8523.4774    241.357     35.315      0.000    8050.390    8
996.565
Airport           2719.5114    302.383      8.994      0.000    2126.805    3
312.217
Commercial_Farm  -3682.1618    459.439     -8.014      0.000   -4582.714   -2
781.609
yr_built         -1212.6579     43.123    -28.121      0.000   -1297.184   -1
128.132
view              2.325e+04   1406.719     16.525      0.000    2.05e+04
2.6e+04
Educational      -1819.6815    795.208     -2.288      0.022   -3378.381    -
260.982
Abandoned        -4935.2213    367.285    -13.437      0.000   -5655.142   -4
215.301
bedrooms          9166.1534   1087.440      8.429      0.000    7034.647
1.13e+04
floorsx2          1.169e+04   1087.464     10.754      0.000    9563.187
1.38e+04
Police            2211.3041    254.203      8.699      0.000    1713.038    2
709.570
waterfront        1.907e+05   1.62e+04     11.769      0.000    1.59e+05
2.23e+05
Government       -8779.7944   1124.473     -7.808      0.000    -1.1e+04   -6
575.699
sqft_living15       86.3963      2.017     42.829      0.000      82.442
90.350
Fire             -1625.7883    692.757     -2.347      0.019   -2983.672    -
267.905
bathroomsx4       9538.8444    457.772     20.838      0.000    8641.559
1.04e+04
Cultural          7110.9110    644.026     11.041      0.000    5848.547    8
373.275
Public_Gathering  3767.3616   1545.292      2.438      0.015     738.414    6
796.310
condition         2.493e+04   1323.502     18.833      0.000    2.23e+04
2.75e+04
zipcode           -129.2149     19.937     -6.481      0.000    -168.293
```

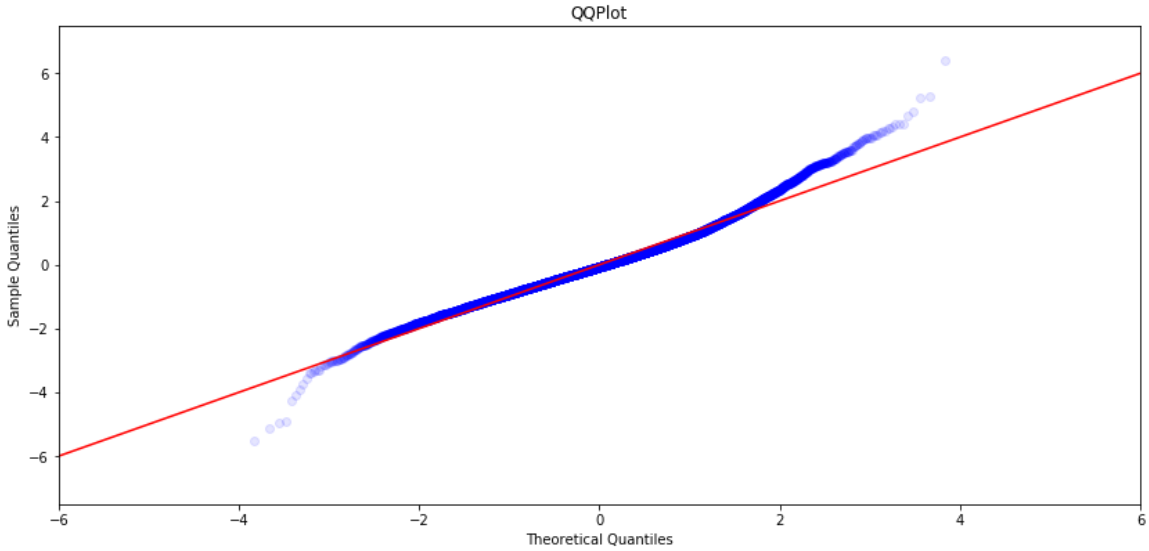
```
-90.136
date                87.3496      6.915     12.632      0.000      73.796
100.904
==============================================================================
=
Omnibus:                     1022.152   Durbin-Watson:                   2.01
6
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1723.59
1
Skew:                           0.505   Prob(JB):                         0.0
0
Kurtosis:                       4.263   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.
Number of features: 29
sqft_lot15
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  price   R-squared:                       0.75
0
Model:                            OLS   Adj. R-squared:                  0.74
9
Method:                 Least Squares   F-statistic:                     163
0.
Date:                Tue, 01 Dec 2020   Prob (F-statistic):               0.0
0
Time:                        10:52:30   Log-Likelihood:            -2.0428e+0
5
No. Observations:               15824   AIC:                         4.086e+0
5
Df Residuals:                   15794   BIC:                         4.088e+0
5
Df Model:                          29
Covariance Type:            nonrobust
==============================================================================
=======
                       coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------------------
-------
Intercept        -7.226e+06   2.07e+06     -3.490      0.000   -1.13e+07   -
3.17e+06
sqft_lot             0.3033      0.038      7.879      0.000       0.228
0.379
long              6.562e+04   1.01e+04      6.469      0.000    4.57e+04
8.55e+04
lat               6.315e+05   9004.631     70.135      0.000    6.14e+05
6.49e+05
```

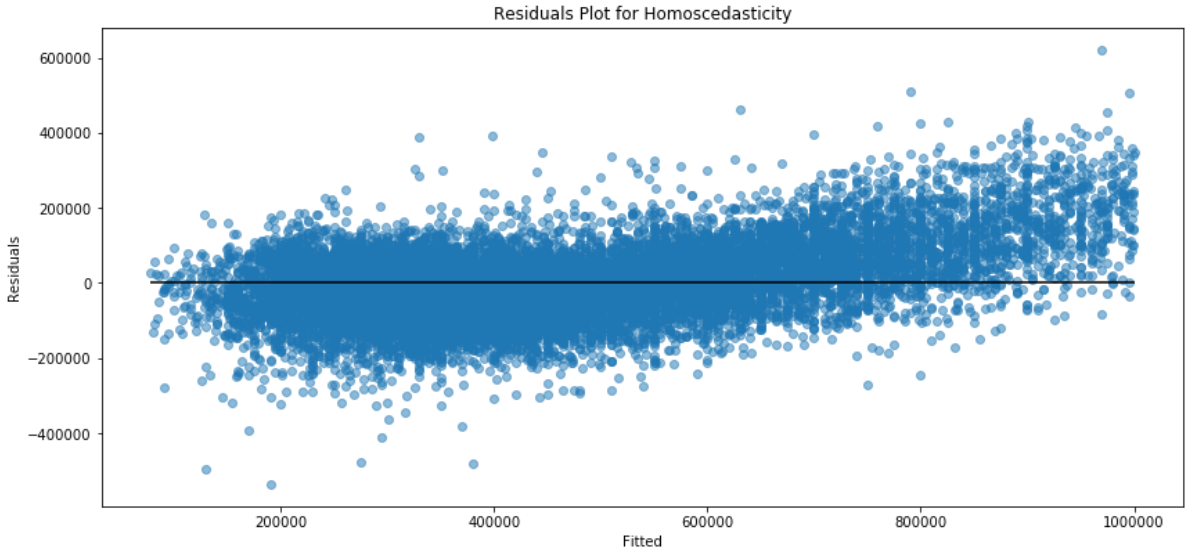
```
sqft_basement       13.8590      2.473      5.604      0.000       9.011
18.707
Lodging           6520.7592    656.695      9.930      0.000    5233.563    7
807.956
Campground        3401.5389    268.697     12.659      0.000    2874.863    3
928.215
Access_Point     -6493.8412    288.678    -22.495      0.000   -7059.682   -5
928.000
grade             6.772e+04   1233.681     54.895      0.000    6.53e+04
7.01e+04
Gated_w_Building  8486.3754    241.542     35.134      0.000    8012.926    8
959.825
Airport           2695.4992    302.599      8.908      0.000    2102.370    3
288.629
Commercial_Farm  -3766.6186    459.558     -8.196      0.000   -4667.406   -2
865.832
sqft_lot15        7079.3160   1870.282      3.785      0.000    3413.349
1.07e+04
yr_built         -1230.4151     43.462    -28.310      0.000   -1315.605   -1
145.226
view              2.333e+04   1407.797     16.573      0.000    2.06e+04
2.61e+04
Educational      -1681.5454    796.057     -2.112      0.035   -3241.908    -
121.183
Abandoned        -4806.5907    372.879    -12.890      0.000   -5537.476   -4
075.705
bedrooms          9290.1464   1094.125      8.491      0.000    7145.537
1.14e+04
floorsx2          1.154e+04   1127.387     10.236      0.000    9329.831
1.37e+04
Police            2224.0479    254.574      8.736      0.000    1725.054    2
723.042
waterfront        1.907e+05   1.62e+04     11.754      0.000    1.59e+05
2.23e+05
Government       -8459.9787   1127.706     -7.502      0.000   -1.07e+04   -6
249.546
sqft_living15       86.8984      2.058     42.216      0.000      82.864
90.933
Fire             -1586.0181    694.162     -2.285      0.022   -2946.655    -
225.381
bathroomsx4       9484.8740    458.331     20.694      0.000    8586.493
1.04e+04
Cultural          6965.2476    644.004     10.816      0.000    5702.927    8
227.568
Public_Gathering  4435.3924   1548.871      2.864      0.004    1399.429    7
471.355
condition         2.488e+04   1326.406     18.761      0.000    2.23e+04
2.75e+04
zipcode           -131.7625     19.950     -6.605      0.000    -170.867
-92.658
date                86.5177      6.923     12.497      0.000      72.948
100.087
==============================================================================
=
Omnibus:                     1044.829   Durbin-Watson:                   2.01
6
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1789.59
```

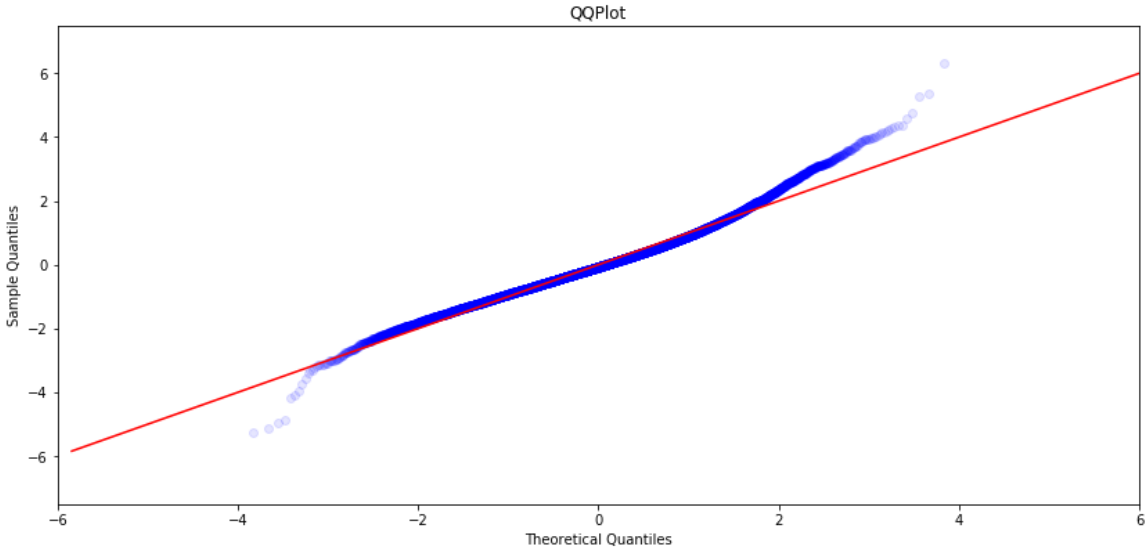
```
4
Skew:                          0.509   Prob(JB):                         0.0
0
Kurtosis:                      4.296   Cond. No.                     2.63e+0
8
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
[2] The condition number is large, 2.63e+08. This might indicate that there a
re
strong multicollinearity or other numerical problems.
```

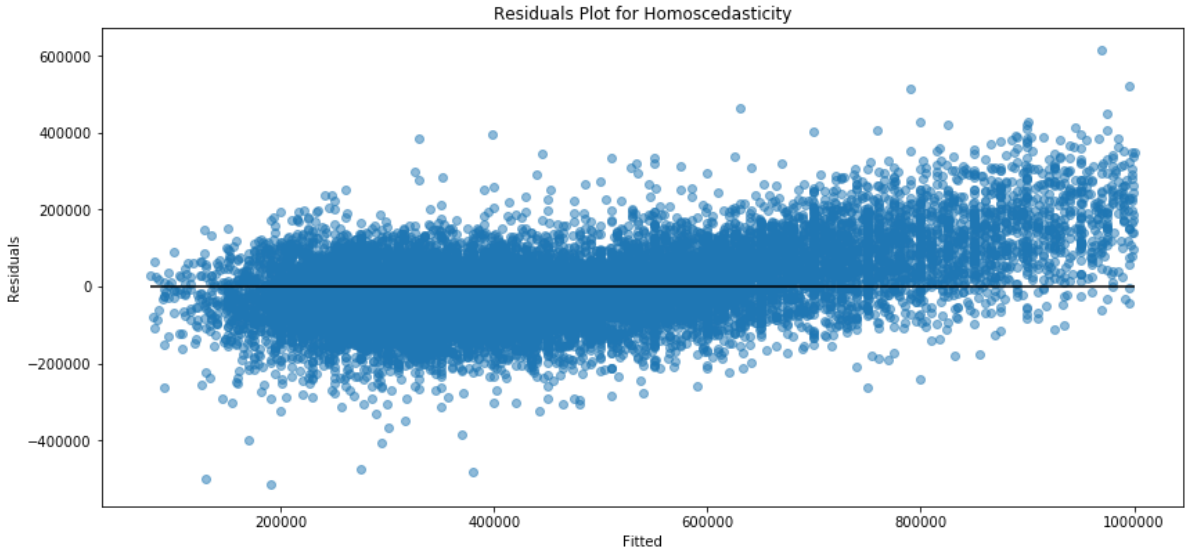


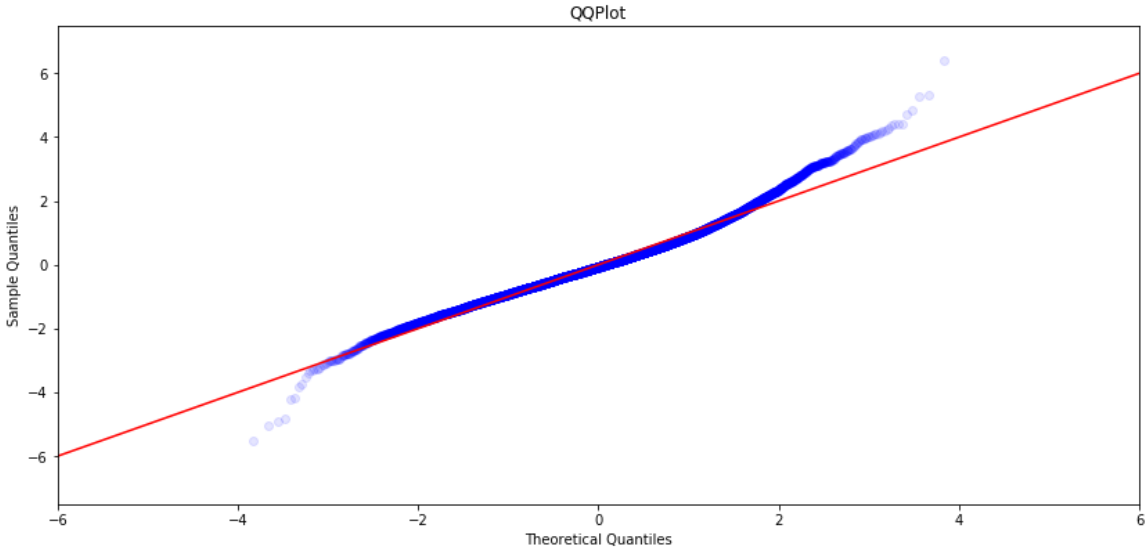


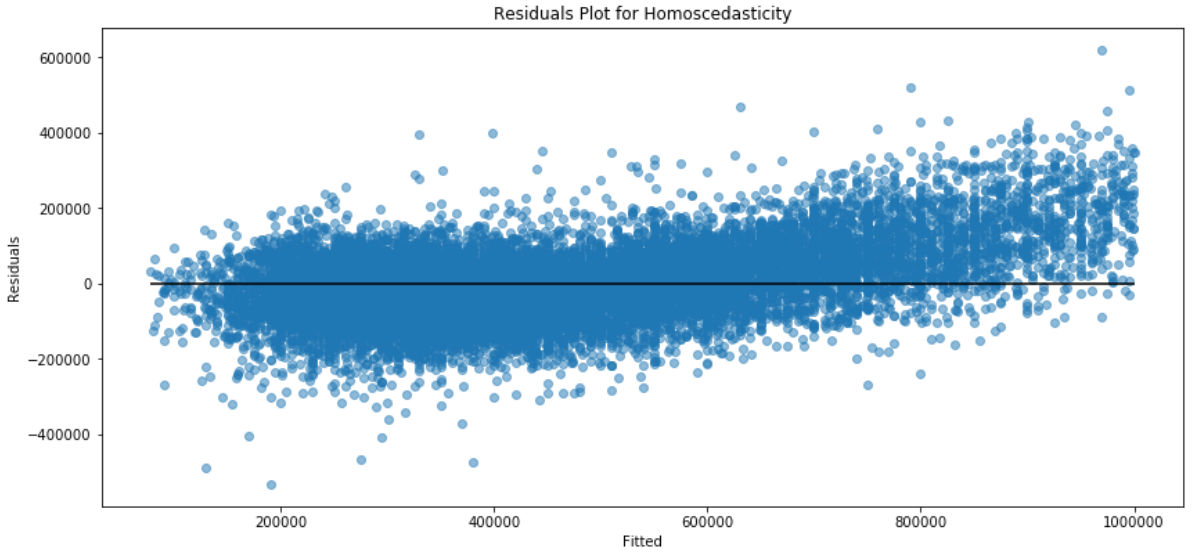

QQPlot
Sample Quantiles
Theoretical Quantiles
6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6

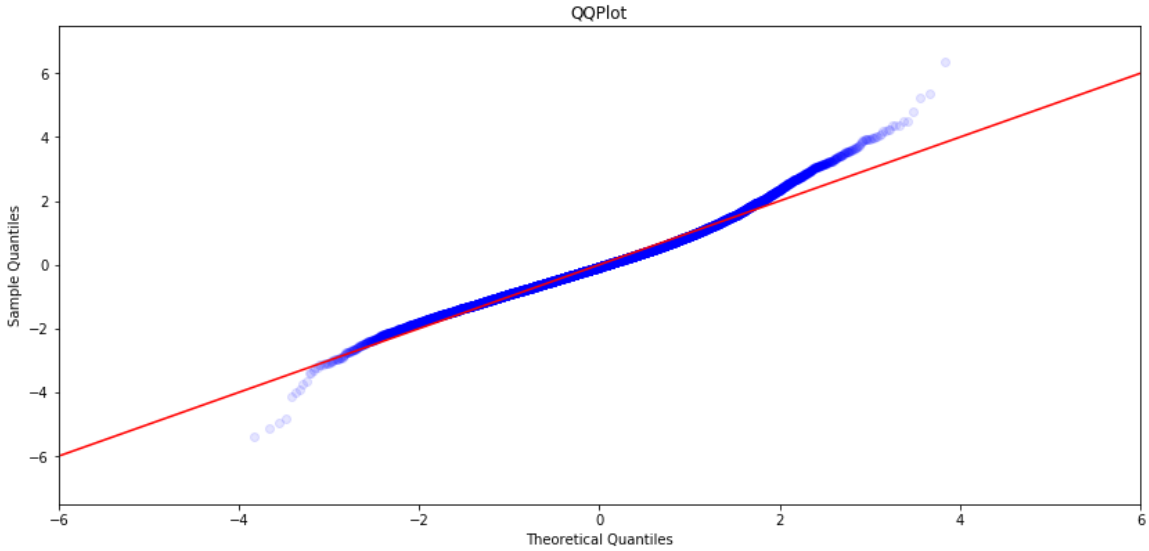

Residuals Plot for Homoscedasticity
Residuals
Fitted
600000
400000
200000
0
−200000
−400000
200000
400000
600000
800000
1000000

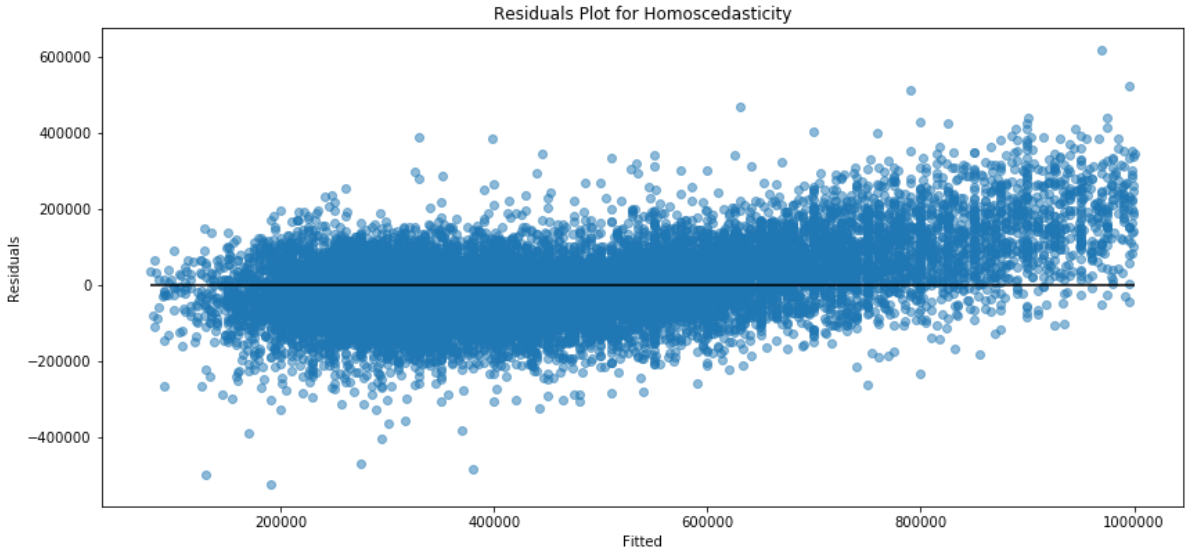

QQPlot
Sample Quantiles
Theoretical Quantiles
6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6

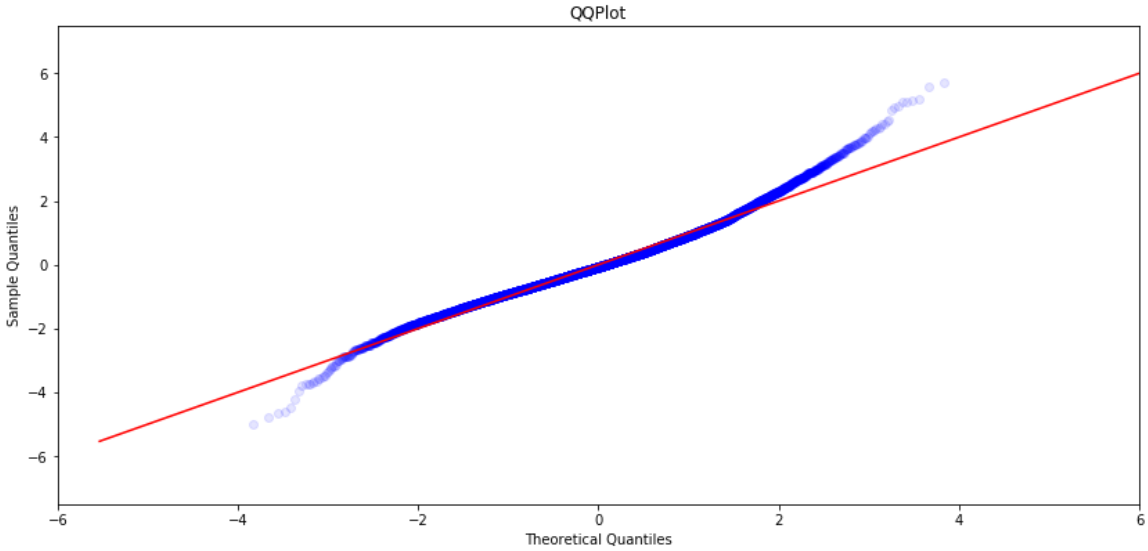

Residuals Plot for Homoscedasticity
Residuals
Fitted
600000
400000
200000
0
−200000
−400000
200000
400000
600000
800000
1000000

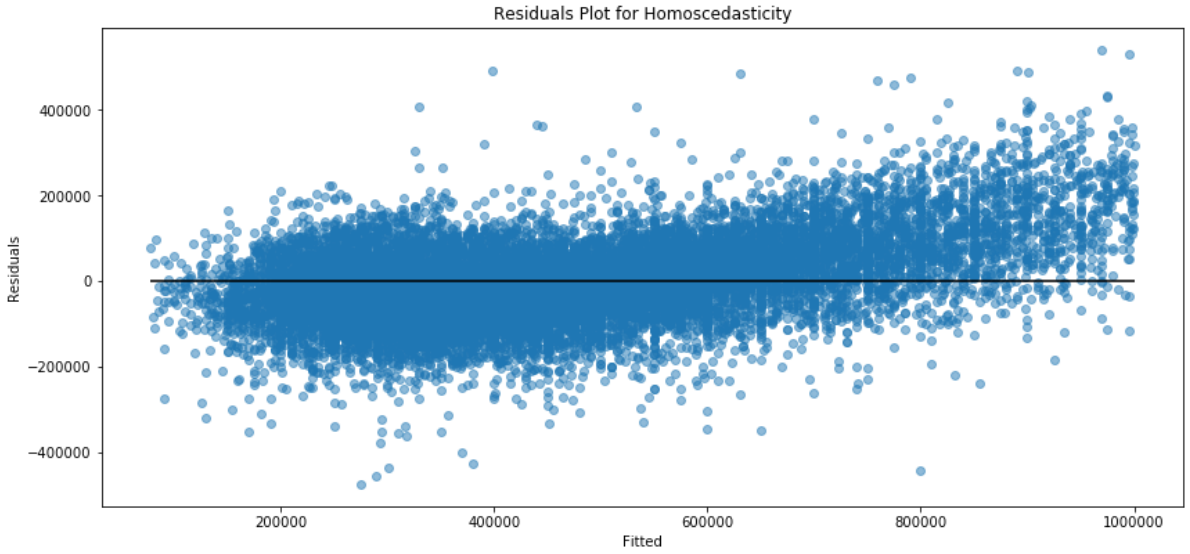

QQPlot
Sample Quantiles
Theoretical Quantiles
6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6

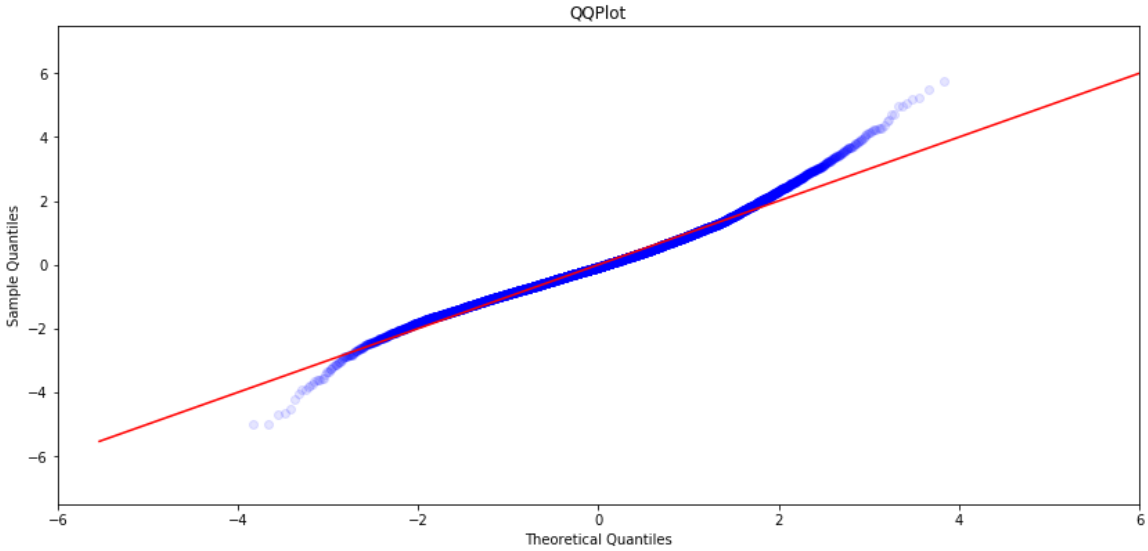

Residuals Plot for Homoscedasticity
Residuals
Fitted
600000
400000
200000
0
−200000
−400000
200000
400000
600000
800000
1000000

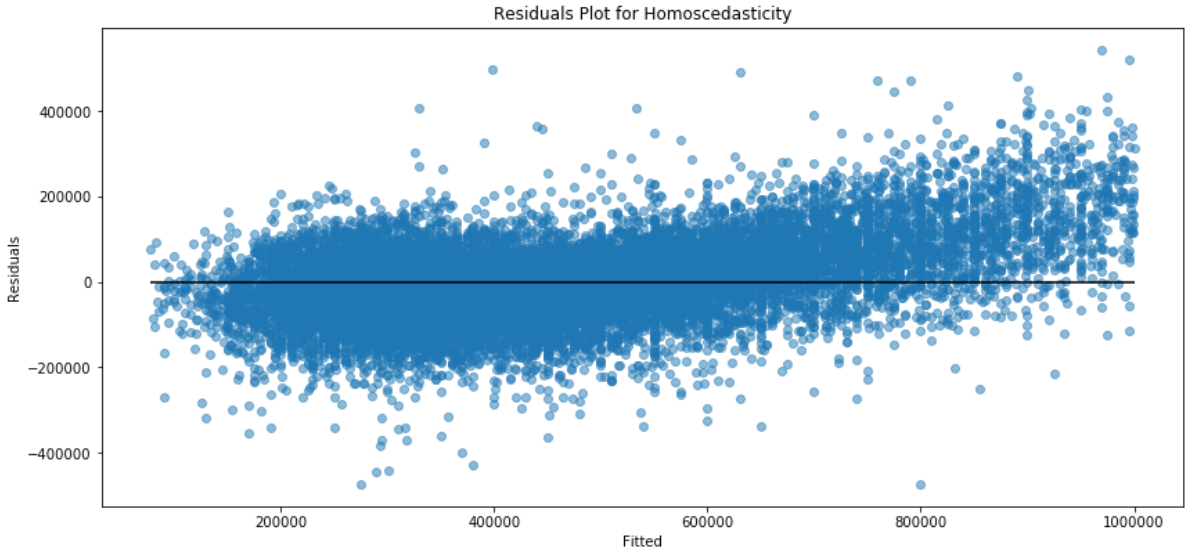

QQPlot
Sample Quantiles
Theoretical Quantiles
6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6

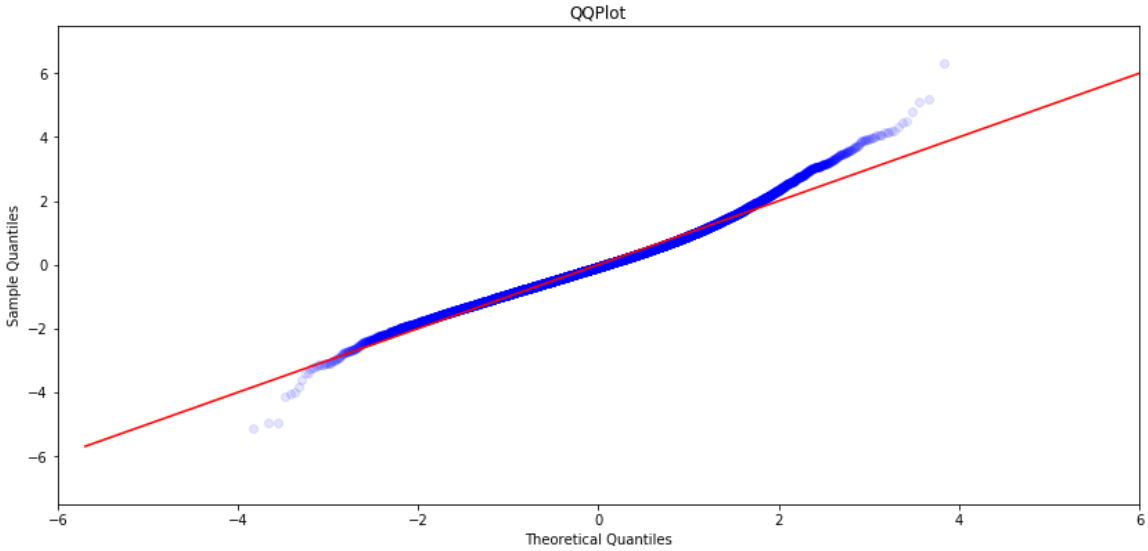

Residuals Plot for Homoscedasticity
Residuals
Fitted
600000
400000
200000
0
−200000
−400000
200000
400000
600000
800000
1000000

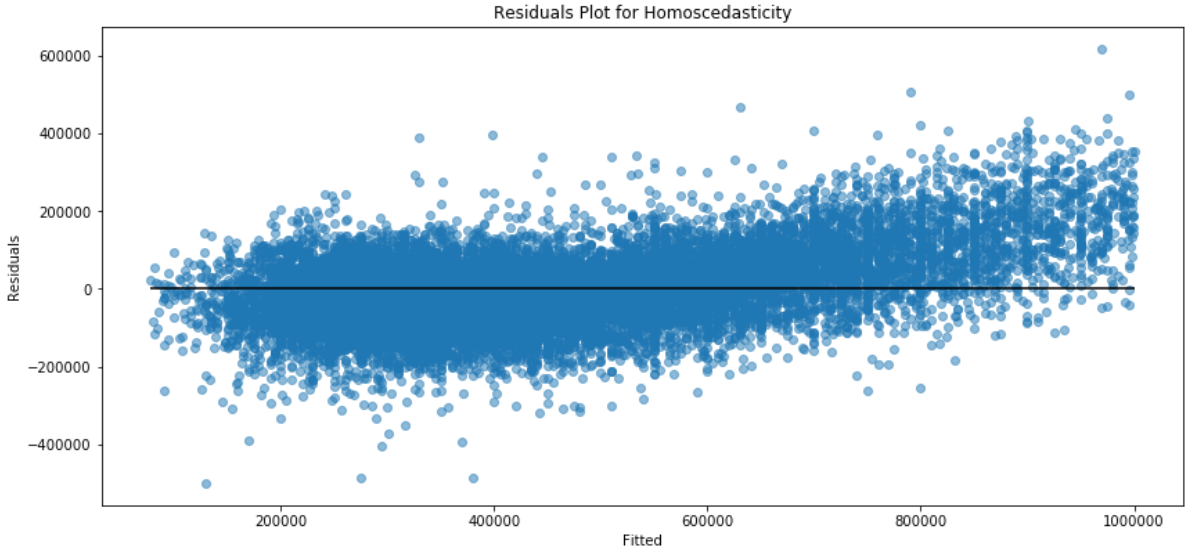

QQPlot
Sample Quantiles
Theoretical Quantiles
6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6

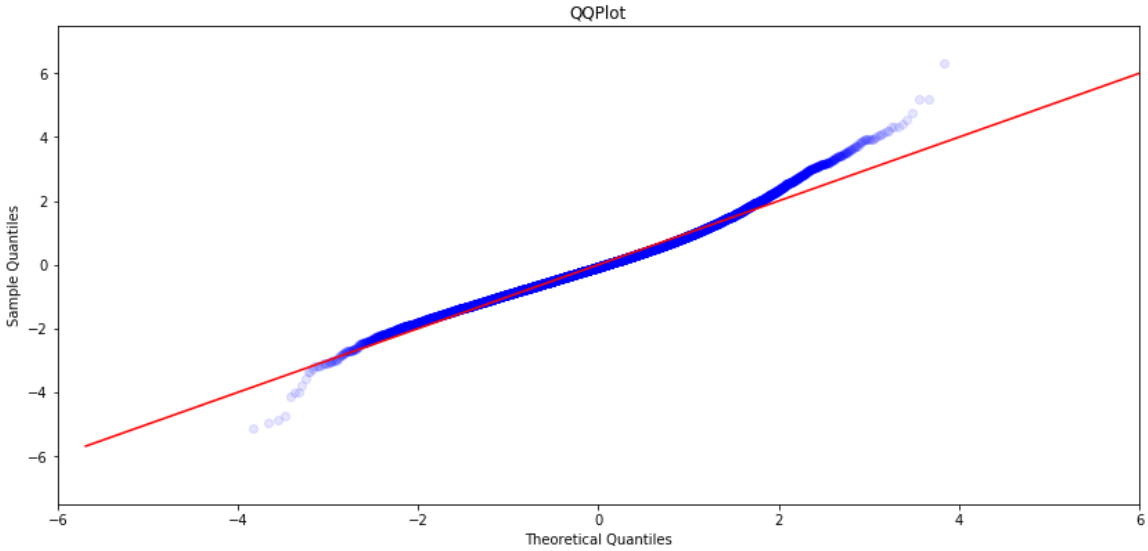

Residuals Plot for Homoscedasticity
Residuals
Fitted
400000
200000
0
−200000
−400000
200000
400000
600000
800000
1000000

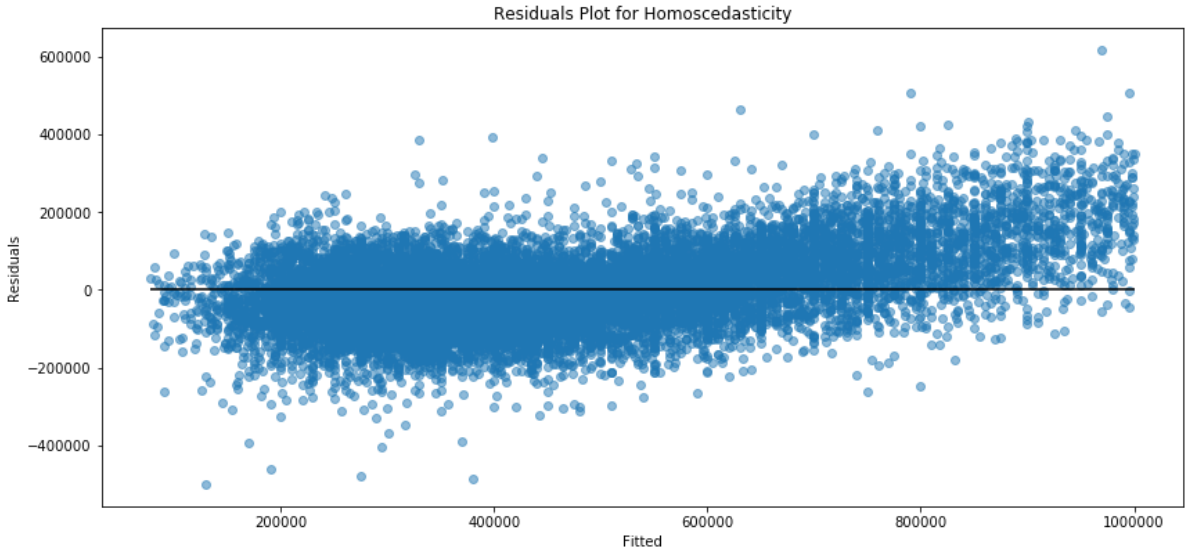

QQPlot
Sample Quantiles
Theoretical Quantiles
6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6


Residuals Plot for Homoscedasticity
Residuals
Fitted
400000
200000
0
−200000
−400000
200000
400000
600000
800000
1000000


QQPlot
Sample Quantiles
Theoretical Quantiles
6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6


Residuals Plot for Homoscedasticity
Residuals
Fitted
600000
400000
200000
0
−200000
−400000
200000
400000
600000
800000
1000000





Benefitting from being log transformed are:

- Gate_wo_building
- Police
- Seasonal Home
- price
- sqft_lot

Benefitting from being scaled and then log transformed are:

- sqft_above
- sqft_living

In [25]:

```
# Model so far:
multicollinearity_threshold=0.7
alpha=0.1
to_minmax = ['sqft_above', 'sqft_living']
to_log = ['Gate_wo_Building', 'Police', 'Seasonal_Home', 'price', 'sqft_lot']
to_ohe = ['lat_long']

data_t = data_train.copy()
data_t = bin_basement(data_t)
data_t = minmax_plus(data_t, to_minmax)
data_t = log(data_t, to_log)
data_t = bin_latlong(data_t)
data_t = ohe(data_t, to_ohe)

x_cols = data_t.drop([outcome], axis=1).columns
x_cols = multicoll_remove(data_t, x_cols, multicollinearity_threshold)
x_cols = simple_selector(data_t, x_cols)

results = model(data_t, x_cols)
metrics(data_t, results, x_cols)
results.summary()
```

```
Number of features: 33
```

Out[25]:

OLS Regression Results

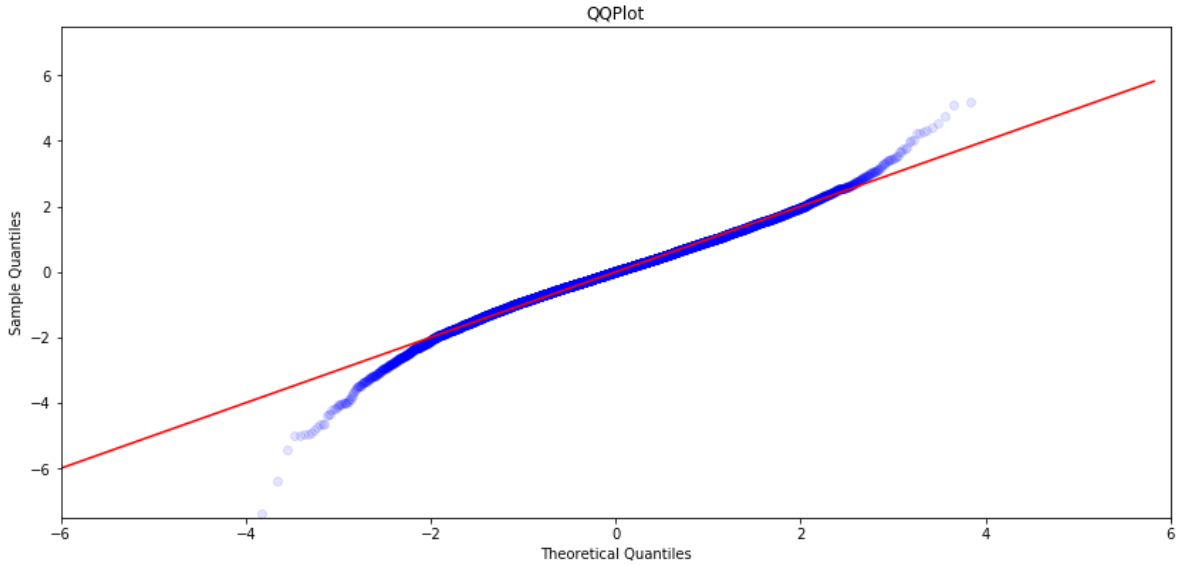
Dep. Variable:	price	**R-squared:**	0.771
Model:	OLS	**Adj. R-squared:**	0.771
Method:	Least Squares	**F-statistic:**	1613.
Date:	Tue, 01 Dec 2020	**Prob (F-statistic):**	0.00
Time:	10:52:54	**Log-Likelihood:**	2402.8
No. Observations:	15824	**AIC:**	-4738.
Df Residuals:	15790	**BIC:**	-4477.
Df Model:	33		
Covariance Type:	nonrobust		

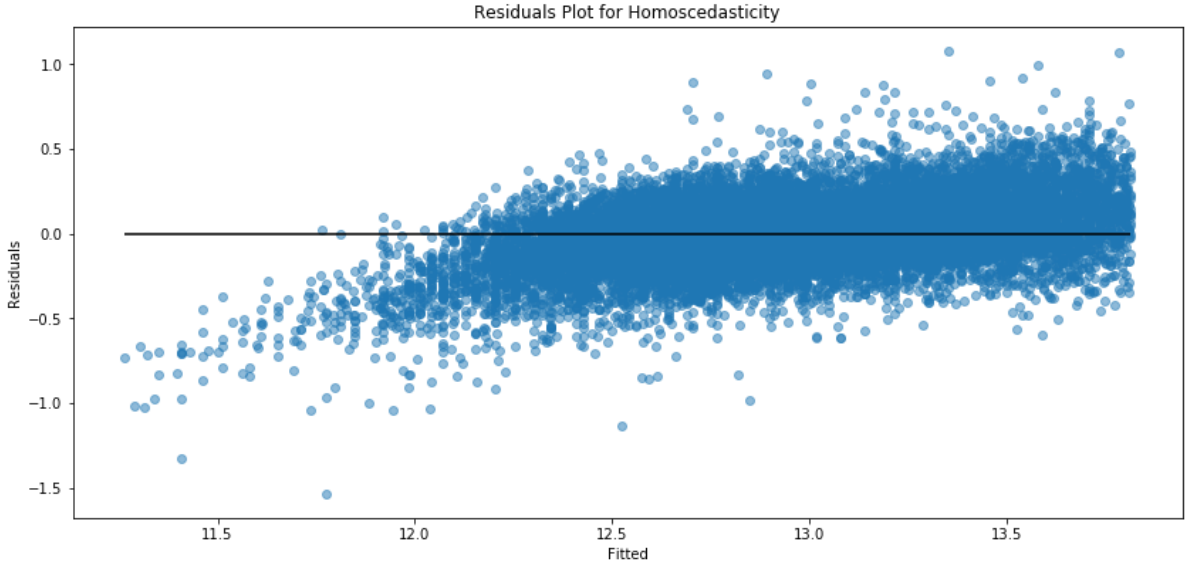
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.1938	4.385	0.272	0.785	-7.402	9.790
sqft_lot	0.0455	0.003	14.942	0.000	0.040	0.052
long	0.1996	0.021	9.352	0.000	0.158	0.241
lat_long_28	-0.7606	0.084	-9.038	0.000	-0.926	-0.596
lat	1.5960	0.019	82.731	0.000	1.558	1.634
sqft_basement	-0.0393	0.004	-9.448	0.000	-0.048	-0.031
Lodging	0.0163	0.001	11.799	0.000	0.014	0.019
Campground	0.0097	0.001	16.969	0.000	0.009	0.011
Access_Point	-0.0111	0.001	-18.220	0.000	-0.012	-0.010
grade	0.1350	0.003	51.367	0.000	0.130	0.140
Gated_w_Building	0.0215	0.001	40.449	0.000	0.020	0.023
Airport	0.0020	0.001	3.103	0.002	0.001	0.003
Commercial_Farm	-0.0052	0.001	-5.318	0.000	-0.007	-0.003
lat_long_20	-0.5563	0.211	-2.642	0.008	-0.969	-0.144
yr_built	-0.0019	9.38e-05	-20.354	0.000	-0.002	-0.002
lat_long_14	0.1460	0.015	9.966	0.000	0.117	0.175
view	0.0444	0.003	14.898	0.000	0.039	0.050
Abandoned	-0.0118	0.001	-14.918	0.000	-0.013	-0.010
lat_long_27	-0.5710	0.100	-5.716	0.000	-0.767	-0.375
bedrooms	0.0237	0.002	10.328	0.000	0.019	0.028
lat_long_3	-0.2018	0.017	-11.758	0.000	-0.235	-0.168
floorsx2	0.0334	0.002	14.060	0.000	0.029	0.038
Police	0.0587	0.003	18.788	0.000	0.053	0.065
lat_long_25	-1.0237	0.153	-6.682	0.000	-1.324	-0.723
Utility	0.0101	0.004	2.721	0.007	0.003	0.017

waterfront	0.3995	0.034	11.595	0.000	0.332	0.467
Government	-0.0177	0.002	-7.285	0.000	-0.022	-0.013
sqft_living15	0.0002	4.34e-06	37.859	0.000	0.000	0.000
bathroomsx4	0.0227	0.001	23.530	0.000	0.021	0.025
Cultural	0.0182	0.001	13.096	0.000	0.015	0.021
Public_Gathering	-0.0105	0.003	-3.269	0.001	-0.017	-0.004
condition	0.0600	0.003	21.349	0.000	0.055	0.066
zipcode	-0.0004	4.27e-05	-9.220	0.000	-0.000	-0.000
date	0.0002	1.47e-05	13.847	0.000	0.000	0.000

Omnibus:	644.795	**Durbin-Watson:**	2.011
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1646.782
Skew:	-0.207	**Prob(JB):**	0.00
Kurtosis:	4.525	**Cond. No.**	2.60e+08

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.6e+08. This might indicate that there are
strong multicollinearity or other numerical problems.


QQPlot
Sample Quantiles
Theoretical Quantiles
6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6


Residuals Plot for Homoscedasticity
Residuals
Fitted
1.0
0.5
0.0
−0.5
−1.0
−1.5
11.5
12.0
12.5
13.0
13.5

In [26]:

```
# Model so far:
multicollinearity_threshold=0.7
alpha=0.1
to_minmax = ['sqft_above', 'sqft_living']
to_log = ['Gate_wo_Building', 'Police', 'Seasonal_Home', 'price', 'sqft_lot']
to_ohe = ['lat_long']

#Transformation
data_t = data_train.copy()
data_t = bin_basement(data_t)
data_t = minmax_plus(data_t, to_minmax)
data_t = log(data_t, to_log)
data_t = bin_latlong(data_t)
data_t = ohe(data_t, to_ohe)

#Feature selection
x_cols = data_t.drop([outcome], axis=1).columns
x_cols = multicoll_remove(data_t, x_cols, multicollinearity_threshold)
x_cols = simple_selector(data_t, x_cols)

results = model(data_t, x_cols)
metrics(data_t, results, x_cols)
results.summary()
```

```
Number of features: 33
```

Out[26]:

OLS Regression Results

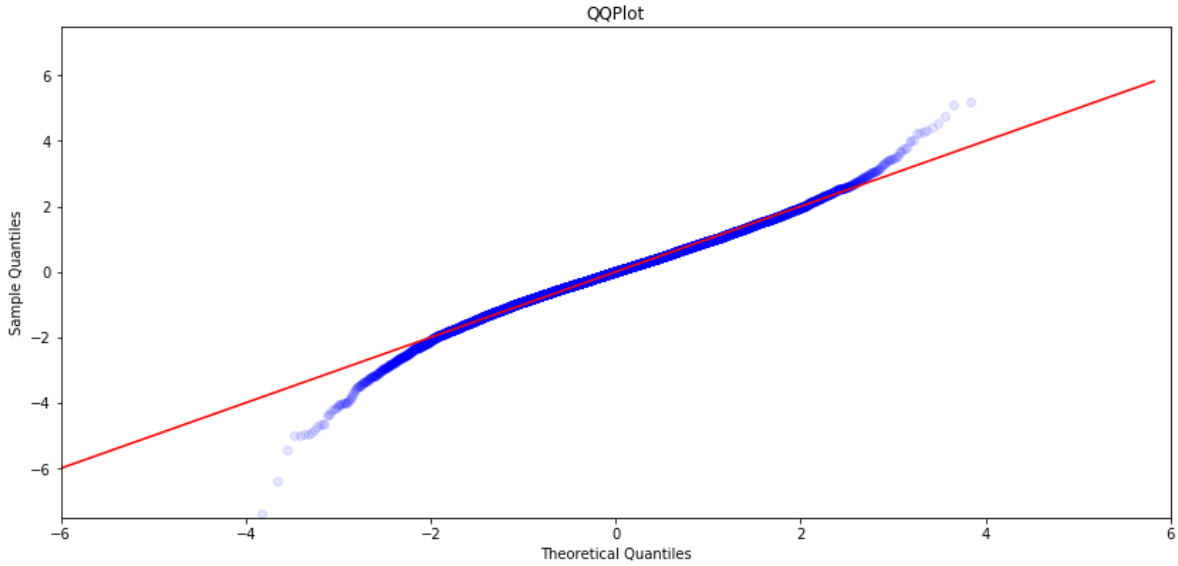
Dep. Variable:	price	**R-squared:**	0.771
Model:	OLS	**Adj. R-squared:**	0.771
Method:	Least Squares	**F-statistic:**	1613.
Date:	Tue, 01 Dec 2020	**Prob (F-statistic):**	0.00
Time:	10:53:00	**Log-Likelihood:**	2402.8
No. Observations:	15824	**AIC:**	-4738.
Df Residuals:	15790	**BIC:**	-4477.
Df Model:	33		
Covariance Type:	nonrobust		

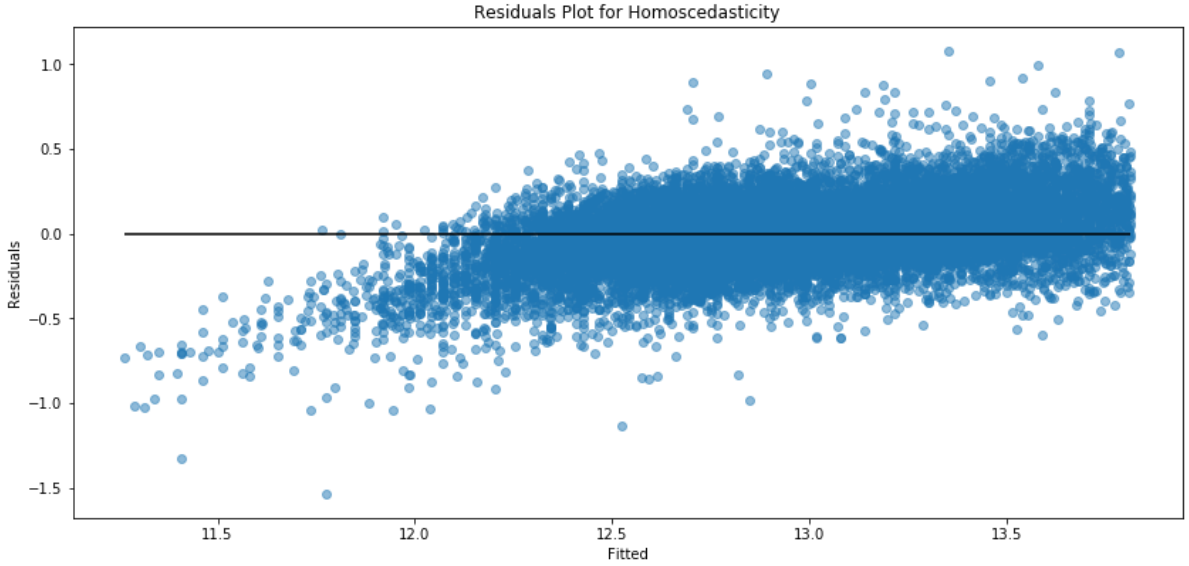
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.1938	4.385	0.272	0.785	-7.402	9.790
sqft_lot	0.0455	0.003	14.942	0.000	0.040	0.052
long	0.1996	0.021	9.352	0.000	0.158	0.241
lat_long_28	-0.7606	0.084	-9.038	0.000	-0.926	-0.596
lat	1.5960	0.019	82.731	0.000	1.558	1.634
sqft_basement	-0.0393	0.004	-9.448	0.000	-0.048	-0.031
Lodging	0.0163	0.001	11.799	0.000	0.014	0.019
Campground	0.0097	0.001	16.969	0.000	0.009	0.011
Access_Point	-0.0111	0.001	-18.220	0.000	-0.012	-0.010
grade	0.1350	0.003	51.367	0.000	0.130	0.140
Gated_w_Building	0.0215	0.001	40.449	0.000	0.020	0.023
Airport	0.0020	0.001	3.103	0.002	0.001	0.003
Commercial_Farm	-0.0052	0.001	-5.318	0.000	-0.007	-0.003
lat_long_20	-0.5563	0.211	-2.642	0.008	-0.969	-0.144
yr_built	-0.0019	9.38e-05	-20.354	0.000	-0.002	-0.002
lat_long_14	0.1460	0.015	9.966	0.000	0.117	0.175
view	0.0444	0.003	14.898	0.000	0.039	0.050
Abandoned	-0.0118	0.001	-14.918	0.000	-0.013	-0.010
lat_long_27	-0.5710	0.100	-5.716	0.000	-0.767	-0.375
bedrooms	0.0237	0.002	10.328	0.000	0.019	0.028
lat_long_3	-0.2018	0.017	-11.758	0.000	-0.235	-0.168
floorsx2	0.0334	0.002	14.060	0.000	0.029	0.038
Police	0.0587	0.003	18.788	0.000	0.053	0.065
lat_long_25	-1.0237	0.153	-6.682	0.000	-1.324	-0.723
Utility	0.0101	0.004	2.721	0.007	0.003	0.017

waterfront	0.3995	0.034	11.595	0.000	0.332	0.467
Government	-0.0177	0.002	-7.285	0.000	-0.022	-0.013
sqft_living15	0.0002	4.34e-06	37.859	0.000	0.000	0.000
bathroomsx4	0.0227	0.001	23.530	0.000	0.021	0.025
Cultural	0.0182	0.001	13.096	0.000	0.015	0.021
Public_Gathering	-0.0105	0.003	-3.269	0.001	-0.017	-0.004
condition	0.0600	0.003	21.349	0.000	0.055	0.066
zipcode	-0.0004	4.27e-05	-9.220	0.000	-0.000	-0.000
date	0.0002	1.47e-05	13.847	0.000	0.000	0.000

Omnibus:	644.795	**Durbin-Watson:**	2.011
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1646.782
Skew:	-0.207	**Prob(JB):**	0.00
Kurtosis:	4.525	**Cond. No.**	2.60e+08

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.6e+08. This might indicate that there are
strong multicollinearity or other numerical problems.





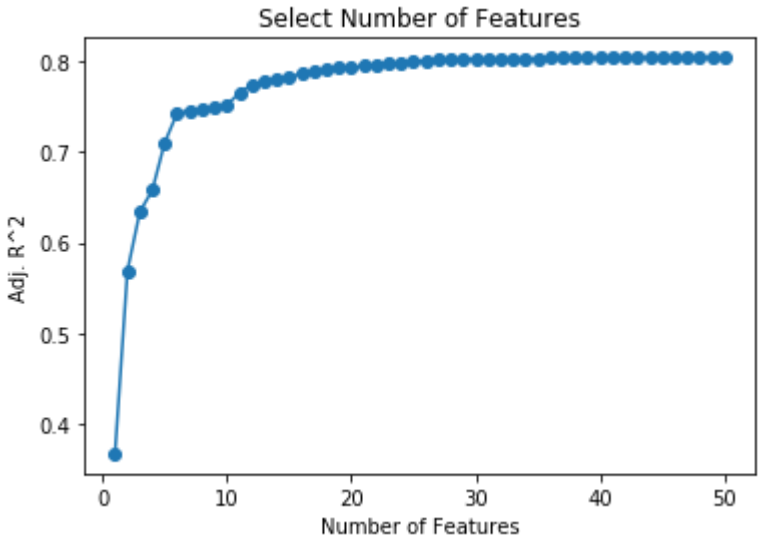
Polynomial Transform and Advanced Feature Selection

Next I tried adding polynomial and interaction terms using the polynomial function. I tried different scaling methods and different orders.
After which I had many more features to narrow down. First I funnelled down the features through gradually stricter simple selection and multicollinear removal. I then pass through stepwise selection and finally through RFE.
I settled on the model you see below:

In [27]:

```
### multicollinearity_threshold=0.7
alpha=0.1
data_t = data_train.copy()

to_minmax = ['sqft_above', 'sqft_living']
to_log = ['Gate_wo_Building', 'Police', 'Seasonal_Home', 'price', 'sqft_lot']
to_ohe = ['lat_long']
to_poly = data_t.drop(['lat', 'long', 'waterfront', 'zipcode',
                       'price'], axis=1).columns
#Transformation
print(data_t.price[0])
data_t = bin_basement(data_t)
data_t = minmax_plus(data_t, to_poly)
print(data_t.price[0])
data_t = log(data_t, to_log)
print(data_t.price[0])
data_t = polynom(data_t, to_poly)
data_t = bin_latlong(data_t)
data_t = ohe(data_t, to_ohe)
x_cols = data_t.drop([outcome], axis=1).columns

#Feature selection
x_cols = simple_selector(data_t, x_cols, alpha=0.15)
x_cols = multicoll_remove(data_t, x_cols, 0.85)
x_cols = simple_selector(data_t, x_cols, alpha=0.1)
x_cols = multicoll_remove(data_t, x_cols, 0.8)
x_cols = stepwise_selector(data_t, x_cols, alpha=0.05)
x_cols = multicoll_remove(data_t, x_cols, 0.7)
data_t = data_t[['price'] + x_cols]
data_t = norm(data_t, x_cols)
x_cols = rfe_selector(data_t, x_cols)


data_t = data_t[['price'] + x_cols]
data_t = data_t.sort_index(axis=1)

print(data_t.price[0])

results = model(data_t, x_cols)
metrics(data_t, results, x_cols)
results.summary()
```

```
410000.0
410000.0
12.92391243868049
86

C:\Users\Maltanno\anaconda3\envs\learn-env\lib\site-packages\numpy\core\fromn
umeric.py:2580: FutureWarning: Method .ptp is deprecated and will be removed
in a future version. Use numpy.ptp instead.
  return ptp(axis=axis, out=out, **kwargs)

85
84
83
82
81
80
79
[(1, 0.36664327006399045), (2, 0.5686740845711282), (3, 0.6340010863654886),
(4, 0.6579308491919927), (5, 0.7105302352901222), (6, 0.7432287483856721),
(7, 0.7456961500932504), (8, 0.7463173693770042), (9, 0.7489009922526184), (1
0, 0.751970362845469), (11, 0.765455144238558), (12, 0.7744104150583462), (1
3, 0.7783508247885412), (14, 0.7807577662771288), (15, 0.7826323528754701),
(16, 0.7870590342980839), (17, 0.7893796251981695), (18, 0.7905119831809312),
(19, 0.7941023770226117), (20, 0.794117285279016), (21, 0.7949435726893987),
(22, 0.7968113129163236), (23, 0.7979222065087225), (24, 0.7991955047070447),
(25, 0.800360020763174), (26, 0.800931337802492), (27, 0.8016895668934931),
(28, 0.801866057314359), (29, 0.8027068444548443), (30, 0.8029506406354711),
(31, 0.80315004381728), (32, 0.8033684337690765), (33, 0.8034067680734317),
(34, 0.8034927667582451), (35, 0.803772358839361), (36, 0.8038912943123171),
(37, 0.8039707017658178), (38, 0.8040270076372182), (39, 0.8040942204720387),
(40, 0.8041309237926598), (41, 0.8041411288396294), (42, 0.8041841636009548),
(43, 0.804286189126233), (44, 0.8042839875517667), (45, 0.8043533558592131),
(46, 0.8043491024196064), (47, 0.8043655459859171), (48, 0.8043592419152787),
(49, 0.804351865111897), (50, 0.8043399788662202)]
```




```
Select number of features
13
12.92391243868049
Number of features: 13
```

Out[27]:

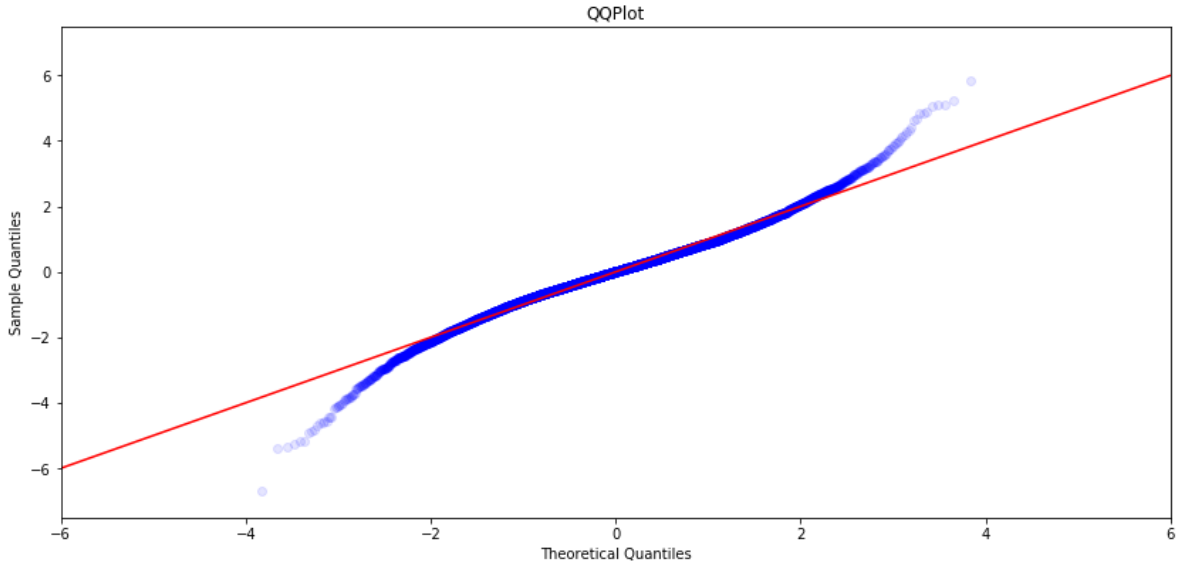
OLS Regression Results

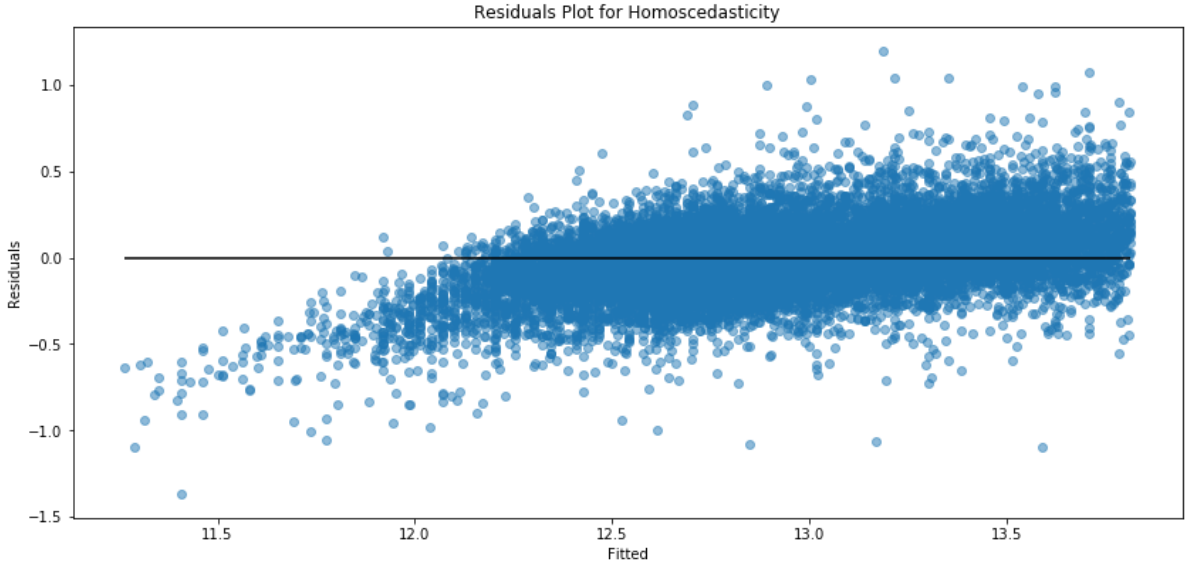
Dep. Variable:	price	**R-squared:**	0.779
Model:	OLS	**Adj. R-squared:**	0.778
Method:	Least Squares	**F-statistic:**	4275.
Date:	Tue, 01 Dec 2020	**Prob (F-statistic):**	0.00
Time:	10:55:42	**Log-Likelihood:**	2661.3
No. Observations:	15824	**AIC:**	-5295.
Df Residuals:	15810	**BIC:**	-5187.
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	12.9606	0.002	7968.338	0.000	12.957	12.964
lat	0.1703	0.003	57.355	0.000	0.164	0.176
Airport__Gate_wo_Building	0.1179	0.005	25.718	0.000	0.109	0.127
grade	0.1130	0.002	48.573	0.000	0.108	0.118
Gate_wo_Building_pow_2	-0.1709	0.004	-42.642	0.000	-0.179	-0.163
Access_Point__view	0.0518	0.002	30.615	0.000	0.048	0.055
Access_Point__Campground	-0.1161	0.003	-41.233	0.000	-0.122	-0.111
Airport__Campground	-0.0951	0.005	-19.631	0.000	-0.105	-0.086
Campground__Seasonal_Home	0.1131	0.006	18.843	0.000	0.101	0.125
Gate_wo_Building__sqft_living	0.1666	0.002	74.943	0.000	0.162	0.171
Cultural_pow_2	0.0330	0.002	16.795	0.000	0.029	0.037
Campground__Gated_w_Building	0.1802	0.003	62.202	0.000	0.175	0.186
condition	0.0417	0.002	24.894	0.000	0.038	0.045
Access_Point__Airport	-0.0563	0.003	-16.786	0.000	-0.063	-0.050

Omnibus:	713.353	**Durbin-Watson:**	2.024
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	2371.296
Skew:	-0.100	**Prob(JB):**	0.00
Kurtosis:	4.886	**Cond. No.**	10.1

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





Test Data

Next I run the test data through the same transformations and use the model to predict prices

In [28]:
```
multicollinearity_threshold=0.7
alpha=0.1
data_tt = data_test.copy()
test = True

to_minmax = ['sqft_above', 'sqft_living']
to_log = ['Gate_wo_Building', 'Police', 'Seasonal_Home', 'price', 'sqft_lot']
to_ohe = ['lat_long']
to_poly = data_tt.drop(['lat', 'long', 'waterfront', 'zipcode', 'price'], axis
=1).columns

print(data_tt.price[0])
data_tt = bin_basement(data_tt)
data_tt = minmax_plus(data_tt, to_minmax, test)
print(data_tt.price[0])
data_tt = log(data_tt, to_log)
print(data_tt.price[0])
data_tt = polynom(data_tt, to_poly)
data_tt = bin_latlong(data_tt, test)
data_tt = ohe(data_tt, to_ohe)


#test set may be missing dummies which the model looks for:
missing = list(set(x_cols) -set(list(data_tt.columns)))
for m in missing:
    data_tt[m] = 0
data_tt = data_tt[['price'] + x_cols]
data_tt = data_tt.sort_index(axis=1)
data_tt = norm(data_tt, x_cols, test)
print(data_tt.price[0])

# results
# metrics(data_tt, results, x_cols)
# results.summary()
```

```
462608.0
462608.0
13.044635322188094
13.044635322188094
```

In [29]:
```
y_hat_train = results.predict(data_t)
rmse_train = MSE(unlog(data_t.price), unlog(y_hat_train))**0.5
y_hat_test = results.predict(data_tt)
rmse_test = MSE(unlog(data_tt.price), unlog(y_hat_test))**0.5
print(f'Train RMSE: {rmse_train} \n Test RMSE: {rmse_test}')
```

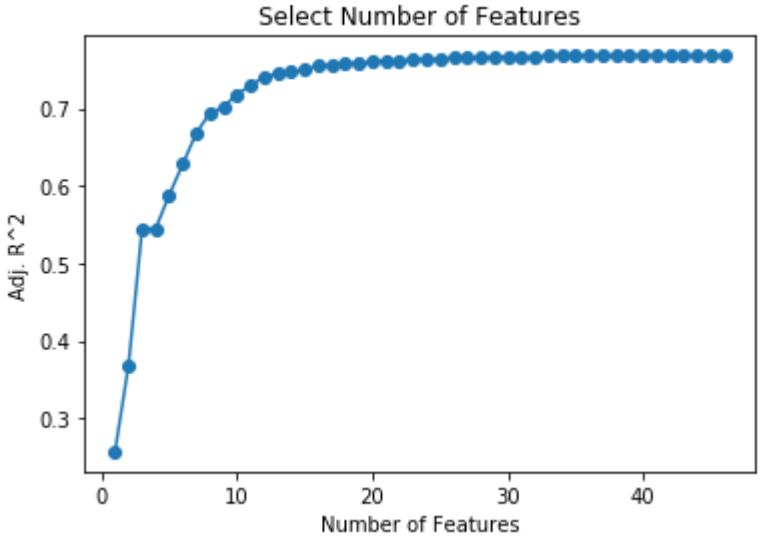
```
Train RMSE: 98204.73789179057
 Test RMSE: 284991.1999283676
```

Train RMSE is ok but Test RMSE is terrible. One thing I wanted to look at was mapping the error, so I decided to try it without any using lat, long, or zipcode:

In [30]:

```
#model with no position:
multicollinearity_threshold=0.7
alpha=0.1
data_t = data_train.copy()

data_t = data_t.drop(['lat', 'long', 'zipcode'], axis=1)

to_minmax = ['sqft_above', 'sqft_living']
to_log = ['Gate_wo_Building', 'Police', 'Seasonal_Home', 'price', 'sqft_lot']
to_poly = data_t.drop(['waterfront',
                       'price'], axis=1).columns

print(data_t.price[0])
data_t = bin_basement(data_t)
data_t = minmax_plus(data_t, to_poly)
print(data_t.price[0])
data_t = log(data_t, to_log)
print(data_t.price[0])
data_t = polynom(data_t, to_poly)
x_cols = data_t.drop([outcome], axis=1).columns

x_cols = simple_selector(data_t, x_cols, alpha=0.15)
x_cols = multicoll_remove(data_t, x_cols, 0.85)
x_cols = simple_selector(data_t, x_cols, alpha=0.1)
x_cols = multicoll_remove(data_t, x_cols, 0.8)
x_cols = stepwise_selector(data_t, x_cols, alpha=0.05)
x_cols = multicoll_remove(data_t, x_cols, 0.7)

data_t = data_t[['price'] + x_cols]
data_t = norm(data_t, x_cols)
x_cols = rfe_selector(data_t, x_cols)


data_t = data_t[['price'] + x_cols]
data_t = data_t.sort_index(axis=1)
print(data_t.price[0])

results = model(data_t, x_cols)
metrics(data_t, results, x_cols)
results.summary()
```

```
410000.0
410000.0
12.92391243868049
81

C:\Users\Maltanno\anaconda3\envs\learn-env\lib\site-packages\numpy\core\fromn
umeric.py:2580: FutureWarning: Method .ptp is deprecated and will be removed
in a future version. Use numpy.ptp instead.
  return ptp(axis=axis, out=out, **kwargs)

80
79
78
77
76
75
74
73
72
71
70
69
[(1, 0.2569593447972993), (2, 0.36838095589019426), (3, 0.5444739413210771),
(4, 0.5447101576390272), (5, 0.5878954833579583), (6, 0.6278038767486371),
(7, 0.667976436129383), (8, 0.6937160662696755), (9, 0.7015804148614089), (1
0, 0.7179741187659758), (11, 0.7293423744908962), (12, 0.740947266935674), (1
3, 0.7444046144017487), (14, 0.747235813866205), (15, 0.749595910067351), (1
6, 0.7547647452498605), (17, 0.7557492317741797), (18, 0.7572268194860452),
(19, 0.7578243045966551), (20, 0.7600540968130227), (21, 0.7613531700434697),
(22, 0.7618075881513494), (23, 0.7627143058672122), (24, 0.7632020399000459),
(25, 0.7637783753877778), (26, 0.7652099575415161), (27, 0.7653640059015249),
(28, 0.7656209089136893), (29, 0.7663371694641887), (30, 0.7666022906970615),
(31, 0.7668995763700552), (32, 0.7670969330505077), (33, 0.76732043271017),
(34, 0.7675127175423846), (35, 0.7676832468510096), (36, 0.7676992921757868),
(37, 0.7677695176400935), (38, 0.7678405903253266), (39, 0.7679220605338718),
(40, 0.7679269428010524), (41, 0.7679216426615307), (42, 0.7679153467483318),
(43, 0.7679064254740824), (44, 0.7678923813745056), (45, 0.7678777892411436),
(46, 0.7678630824151256)]
```



```
Select number of features
12
12.92391243868049
Number of features: 12
```

Out[30]:

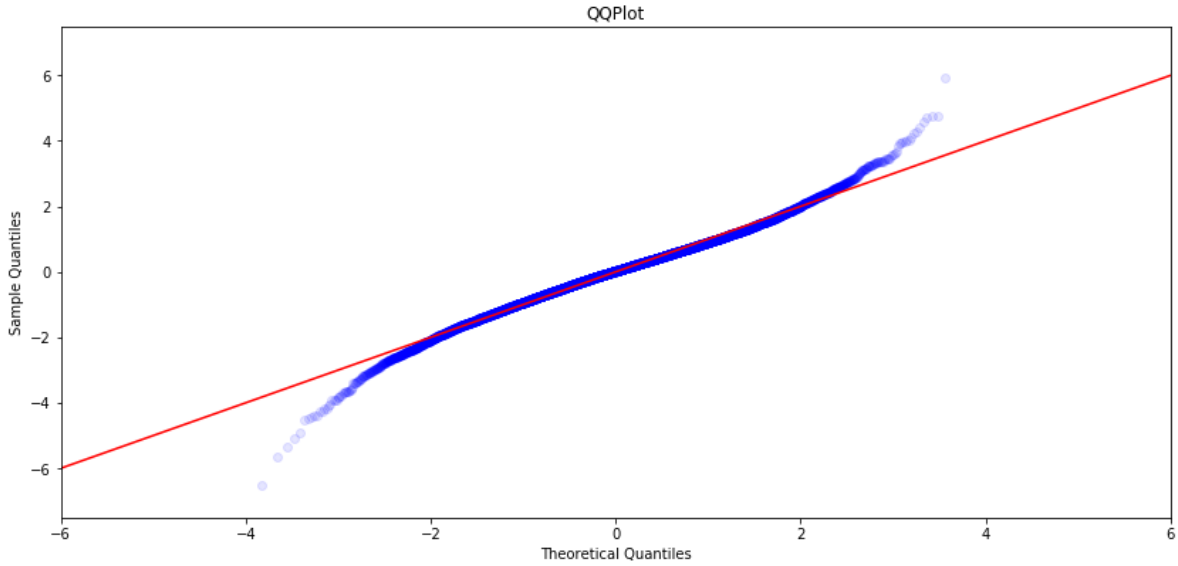
OLS Regression Results

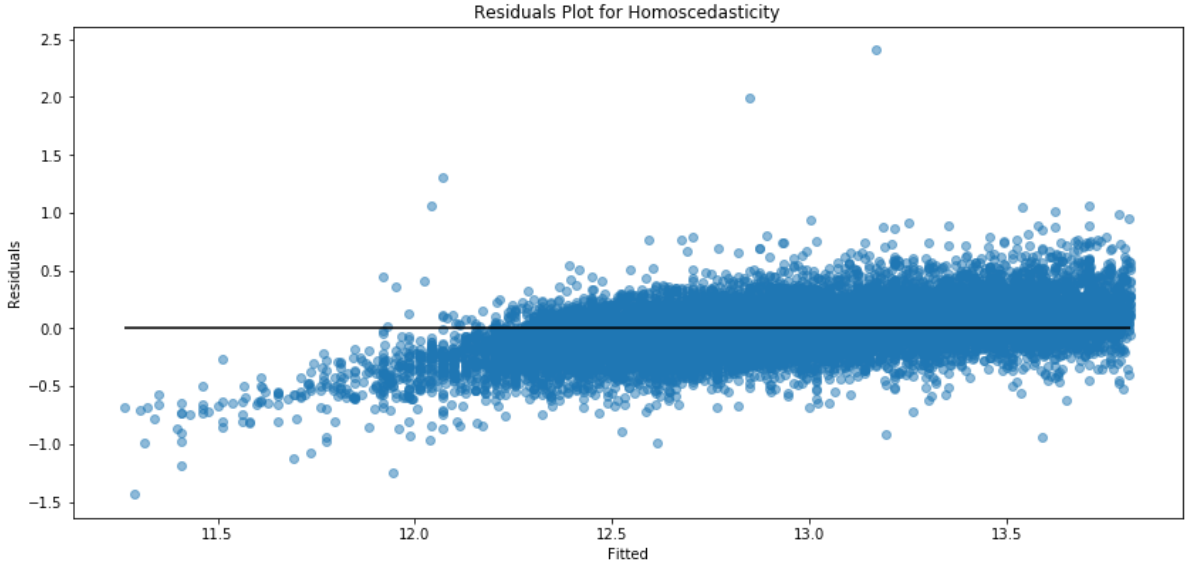
Dep. Variable:	price	**R-squared:**	0.741
Model:	OLS	**Adj. R-squared:**	0.741
Method:	Least Squares	**F-statistic:**	3772.
Date:	Tue, 01 Dec 2020	**Prob (F-statistic):**	0.00
Time:	10:56:57	**Log-Likelihood:**	1427.0
No. Observations:	15824	**AIC:**	-2828.
Df Residuals:	15811	**BIC:**	-2728.
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	12.9606	0.002	7370.665	0.000	12.957	12.964
grade	0.0956	0.002	38.554	0.000	0.091	0.101
Gate_wo_Building_pow_2	-0.3300	0.003	-101.152	0.000	-0.336	-0.324
condition__sqft_living15	0.0811	0.002	36.044	0.000	0.077	0.085
Gated_w_Building__Police	0.0645	0.002	30.062	0.000	0.060	0.069
Access_Point__view	0.0492	0.002	26.633	0.000	0.046	0.053
Campground__Gate_wo_Building	0.4026	0.005	85.362	0.000	0.393	0.412
Access_Point__Campground	-0.0851	0.002	-37.080	0.000	-0.090	-0.081
Airport__Campground	-0.1418	0.005	-28.531	0.000	-0.152	-0.132
Airport__Seasonal_Home	0.1722	0.004	44.103	0.000	0.165	0.180
Gate_wo_Building__sqft_living	0.1381	0.003	53.521	0.000	0.133	0.143
Abandoned__Police	-0.0546	0.002	-25.739	0.000	-0.059	-0.050
Airport__Cemetery	-0.0733	0.003	-23.799	0.000	-0.079	-0.067

Omnibus:	946.460	**Durbin-Watson:**	2.001
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	4227.105
Skew:	0.019	**Prob(JB):**	0.00
Kurtosis:	5.532	**Cond. No.**	7.30

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


QQPlot
Sample Quantiles
Theoretical Quantiles
6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6


Residuals Plot for Homoscedasticity
Residuals
Fitted
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
11.5
12.0
12.5
13.0
13.5

In [31]:
```
multicollinearity_threshold=0.7
alpha=0.1
data_tt = data_test.copy()
test = True

data_tt = data_tt.drop(['lat', 'long', 'zipcode'], axis=1)

to_minmax = ['sqft_above', 'sqft_living']
to_log = ['Gate_wo_Building', 'Police', 'Seasonal_Home', 'price', 'sqft_lot']
to_poly = data_tt.drop(['waterfront',
                        'price'], axis=1).columns

print(data_tt.price[0])
data_tt = bin_basement(data_tt)
data_tt = minmax_plus(data_tt, to_poly)
print(data_tt.price[0])
data_tt = log(data_tt, to_log)
print(data_tt.price[0])
data_tt = polynom(data_tt, to_poly)

#test set may be missing dummies which the model looks for:
missing = list(set(x_cols) -set(list(data_tt.columns)))
for m in missing:
    data_tt[m] = 0
print(data_tt.price[0])
data_tt = data_tt[['price'] + x_cols]
data_tt = data_tt.sort_index(axis=1)
data_tt = norm(data_tt, x_cols, test)
print(data_tt.price[0])
```

```
462608.0
462608.0
13.044635322188094
13.044635322188094
13.044635322188094
```

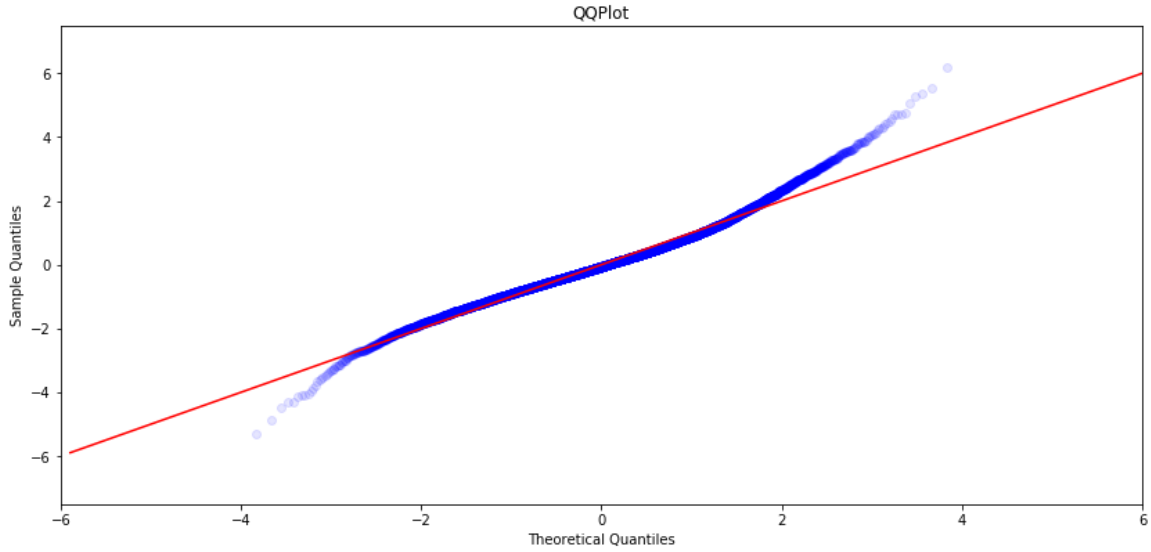
In [32]:
```
y_hat_train = results.predict(data_t)
rmse_train = MSE(unlog(data_t.price), unlog(y_hat_train))**0.5
y_hat_test = results.predict(data_tt)
rmse_test = MSE(unlog(data_tt.price), unlog(y_hat_test))**0.5
print(f'Train RMSE: {rmse_train} \n Test RMSE: {rmse_test}')
```

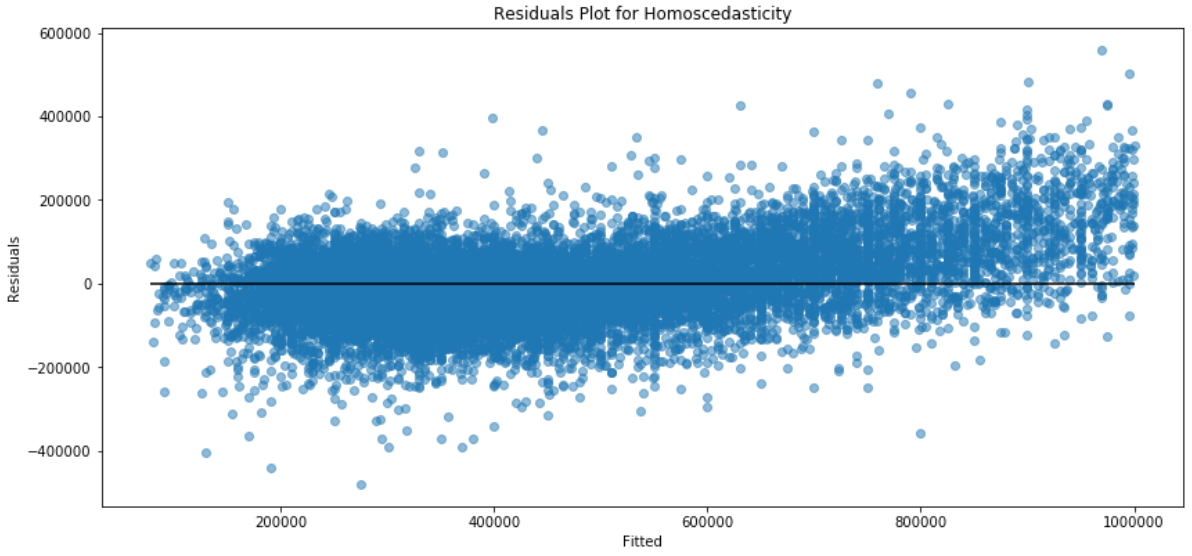
```
Train RMSE: 104126.96359716072
 Test RMSE: 118296.45508329422
```

Although the Train RMSE is a bit worse, the Test RMSE is much better.
We have our final model.
But first let's just check the basemodel's RMSEs:

In [33]:
```
outcome = 'price'
x_cols = data_train.drop([outcome], axis=1).columns
results = model(data_train, x_cols)

metrics(data_train, results, x_cols)
results.summary()

y_hat_train = results.predict(data_train)
rmse_train = MSE(data_train.price, y_hat_train)**0.5
y_hat_test = results.predict(data_test)
rmse_test = MSE(data_test.price, y_hat_test)**0.5
print(f'Train RMSE: {rmse_train} \n Test RMSE: {rmse_test}')
```

```
Number of features: 36
Train RMSE: 90503.58916320122
 Test RMSE: 92445.73893425212
```





With much better RMSEs (and time running out) we're going to go ahead with the base model

Creating an Error Map

First I add the predictions to the dataframe, then calculate the residuals. Then I create 2 dataframes, with just lat, long and residuals; one for the bottom 25% of residuals and one for the top. Next I find the min and max of each for scaling from 0-255. With these I create a colour column. I then use Folium to create a map from these

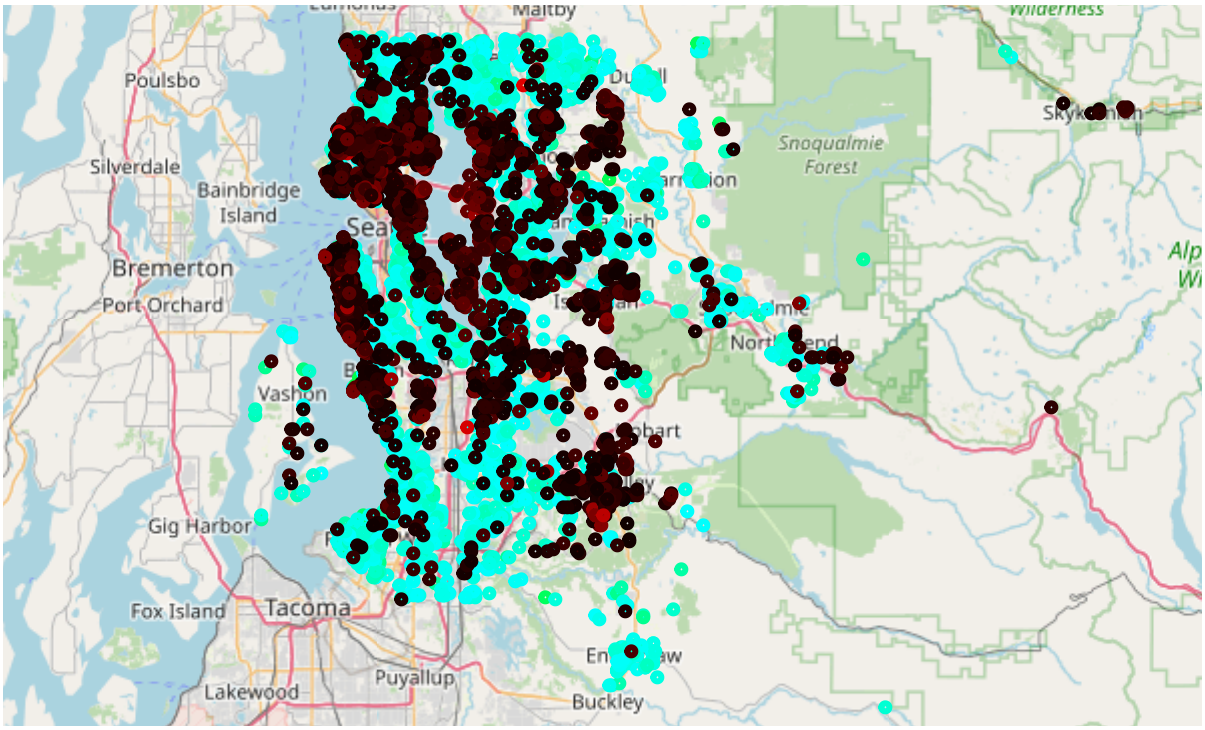
In [34]:
```
data_r = pd.concat([data_train, y_hat_train], axis=1)
data_r.rename(columns={0:'residuals'}, inplace=True)
data_r.residuals = data_r.price - data_r.residuals
data_r  = data_r[['lat', 'long', 'residuals']]
data_ra = data_r[data_r.residuals < data_r.residuals.quantile(0.25)]
data_rb = data_r[data_r.residuals > data_r.residuals.quantile(0.75)]
print(len(data_ra), len(data_rb))
```

```
3956 3956
```

In [35]:
```
minb = data_r.residuals.quantile(0.75)
maxb = data_r.residuals.max()
mina = data_r.residuals.min()
maxa = data_r.residuals.quantile(0.25)


data_rb.loc[:,'colour'] = data_rb.residuals.apply(lambda x:
                          '#'+str(hex(int(255*(x-minb)/(maxb-minb))))[2:]+'0000'
)

data_ra.loc[:,'colour'] = data_ra.residuals.apply(lambda x:
                          '#00ff'+str(hex(int(255*(x-mina)/(maxa-mina))))[2:])

lat = 47.5480
long = -121.9836
base_map = folium.Map([lat,long], zoom_start=9, width=900, height=600)

for df in [data_ra, data_rb]:
    x = df.lat
    y = df.long
    z = df.colour
    points = list(zip(x,y,z))
    for p in points:

        lat = p[0]
        long = p[1]
        marker = folium.CircleMarker(location=[lat,long], radius=2, color=p[2
])
        marker.add_to(base_map)


base_map
```

```
C:\Users\Maltanno\anaconda3\envs\learn-env\lib\site-packages\pandas\core\inde
xing.py:376: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/st
able/user_guide/indexing.html#returning-a-view-versus-a-copy
  self.obj[key] = _infer_fill_value(value)
C:\Users\Maltanno\anaconda3\envs\learn-env\lib\site-packages\pandas\core\inde
xing.py:494: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/st
able/user_guide/indexing.html#returning-a-view-versus-a-copy
  self.obj[item] = s
```

Out[35]:



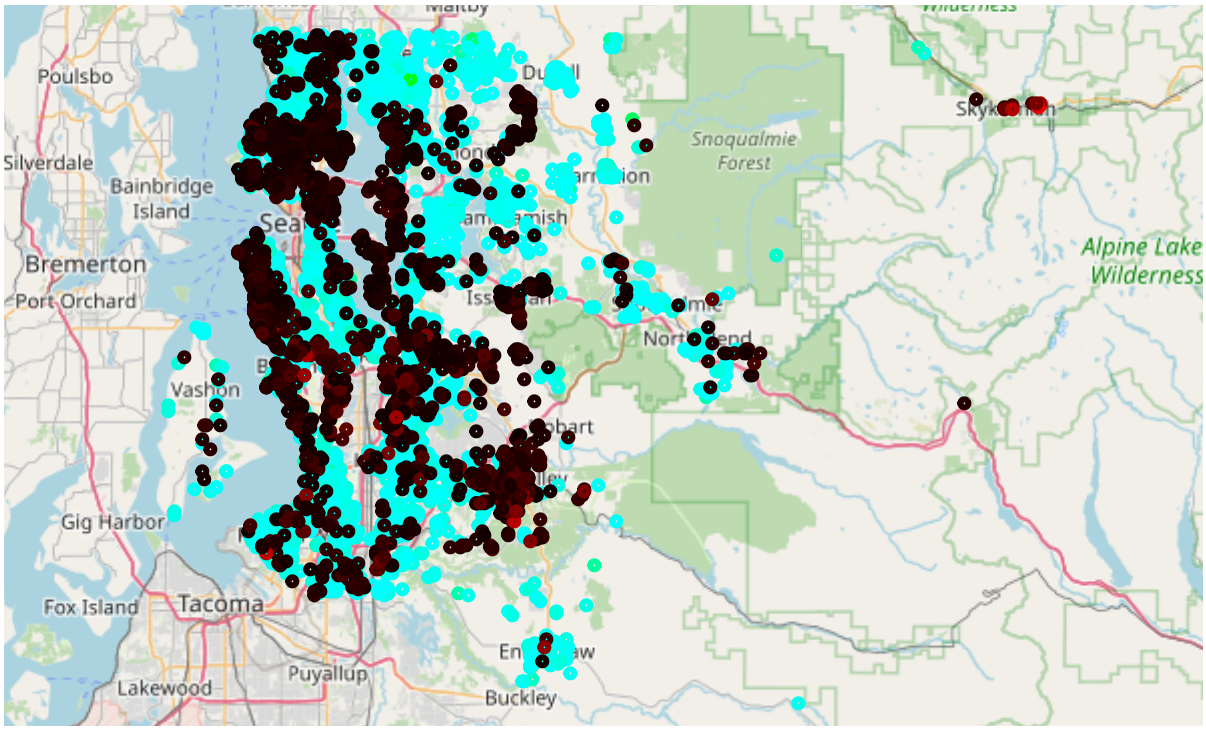
Below I do the same again but this time using error/price. I felt that, otherwise, more expensive places could be overrepresented.

In [36]:
```
data_r = pd.concat([data_train, y_hat_train], axis=1)
data_r.rename(columns={0:'residuals2'}, inplace=True)
data_r.residuals2 = 100*(data_r.price - data_r.residuals2)/data_r.price
data_r  = data_r[['lat', 'long', 'residuals2']]
data_ra = data_r[data_r.residuals2 < data_r.residuals2.quantile(0.25)]
data_rb = data_r[data_r.residuals2 > data_r.residuals2.quantile(0.75)]
print(len(data_ra), len(data_rb))
```

```
3956 3956
```

In [37]:
```
minb = data_r.residuals2.quantile(0.75)
maxb = data_r.residuals2.max()
mina = data_r.residuals2.min()
maxa = data_r.residuals2.quantile(0.25)

data_rb.loc[:,'colour'] = data_rb.residuals2.apply(lambda x:
                          '#'+str(hex(int(255*(x-minb)/(maxb-minb))))[2:]+'0000'
)

data_ra.loc[:,'colour'] = data_ra.residuals2.apply(lambda x:
                          '#00ff'+str(hex(int(255*(x-mina)/(maxa-mina))))[2:])

lat = 47.5480
long = -121.9836
base_map = folium.Map([lat,long], zoom_start=9, width=900, height=600)

for df in [data_ra, data_rb]:
    x = df.lat
    y = df.long
    z = df.colour
    points = list(zip(x,y,z))
    for p in points:

        lat = p[0]
        long = p[1]
        marker = folium.CircleMarker(location=[lat,long], radius=2, color=p[2
])
        marker.add_to(base_map)


base_map
```

Out[37]:



In [ ]:

See README for further results

In [ ]:

In [40]:
```
#Looked briefly at Principal Component Analysis
X = data_train.drop(outcome, axis=1)
pca = PCA(n_components = 4, whiten=True)
pca.fit(X)
```

Out[40]:
```
PCA(copy=True, iterated_power='auto', n_components=4, random_state=None,
    svd_solver='auto', tol=0.0, whiten=True)
```

In [41]:
```
pca.components_
```

Out[41]:
```
array([[ 1.52468036e-05,  2.86686176e-05,  2.13232862e-05,
        -4.57081712e-05,  2.89475648e-05, -2.97217485e-06,
         1.14721067e-05,  8.10096692e-06,  2.37634892e-06,
        -7.14755565e-05, -1.10071879e-05,  8.65422013e-06,
         1.52861316e-05,  1.38293493e-05,  7.94006204e-06,
        -6.94643498e-05,  1.91875051e-06,  6.24200061e-06,
         1.93787646e-06, -5.01264815e-09, -5.68862302e-06,
        -2.51838539e-06,  3.99708883e-06, -3.84474495e-07,
         1.35827512e-06,  4.65912766e-03,  4.86304297e-04,
         5.13205168e-03,  4.15673313e-03,  9.65553091e-01,
         2.60079912e-01,  1.06668393e-06,  1.13582801e-07,
         3.26144027e-05,  2.95071046e-05, -3.01779514e-04],
       [-5.96166034e-05, -1.10047018e-04, -8.74784550e-05,
         1.72362527e-04, -8.48943927e-05,  2.73741666e-06,
        -4.12883680e-05, -3.00445452e-05, -1.35988911e-05,
         2.33211088e-04,  4.64266498e-05, -2.91625288e-05,
        -5.75359797e-05, -3.45994436e-05, -2.80479919e-05,
         2.25470547e-04, -6.15314689e-06, -1.86368262e-05,
        -8.75605269e-06, -3.73958955e-06,  4.63788040e-04,
         1.00004831e-05, -1.55784665e-05,  7.14691081e-07,
        -4.03133658e-06, -1.46553334e-02, -1.70011817e-03,
        -1.63993944e-02, -1.76333612e-02,  2.60203649e-01,
        -9.65139977e-01, -2.76004035e-06, -2.73012860e-07,
        -1.89491314e-04, -1.79433895e-04,  9.38854864e-04],
       [ 1.19296243e-05,  5.18343209e-05,  5.50864859e-04,
        -1.58270204e-03,  5.20114968e-04, -1.67311606e-04,
         3.77494798e-04, -4.79408720e-06,  7.15095358e-05,
        -2.80072436e-03, -7.07623301e-04,  1.30916238e-04,
         2.99577612e-04,  6.73658870e-04,  4.59583264e-05,
        -2.64122289e-03,  2.82039404e-05,  1.84407962e-03,
         4.62711971e-04, -8.80441299e-05, -4.18326710e-03,
         4.58189509e-04,  6.86291233e-04,  1.00143501e-06,
         3.83655461e-05,  6.01519559e-01,  5.17759037e-02,
         6.54238589e-01,  4.54082396e-01, -6.02831012e-04,
        -2.88197220e-02,  6.76575643e-05, -8.42586857e-07,
         1.17883768e-02,  1.17101926e-02, -1.12488444e-02],
       [-2.13512244e-04, -8.95582716e-04, -2.93001576e-04,
         2.07629494e-03, -8.53767977e-04,  3.58408430e-04,
        -4.64963963e-04, -7.84252340e-05,  5.69227164e-05,
         2.79257880e-03,  9.86888620e-04, -1.50874281e-04,
        -4.01861619e-04, -5.11168307e-04, -7.07352093e-05,
         3.53870970e-03,  7.97776330e-06,  6.27716959e-04,
         3.75797273e-04,  2.62525208e-04, -5.21887422e-03,
        -8.25317102e-04, -8.17505185e-05,  2.97342003e-05,
        -6.13493582e-05, -4.60291524e-01,  8.13660418e-01,
         3.54200737e-01,  7.60174007e-03,  5.48310609e-05,
        -5.68162994e-04,  2.18533765e-04,  2.01622215e-06,
        -1.29566191e-02, -1.20039426e-02,  1.42667310e-02]])
```

Future work

- Try some minor changes from the base model e.g. some transforms, without feature selection
- Add cross validation
- Look further into PCA
- Add sklearn metric into RFE function
- Find how to get more data from the API
- If we could get the same data over a longer time, it would be interesting to see how the error map changes
- Try a narrower range of prices