다변량분석 Assignment 3

2014170824 황태민

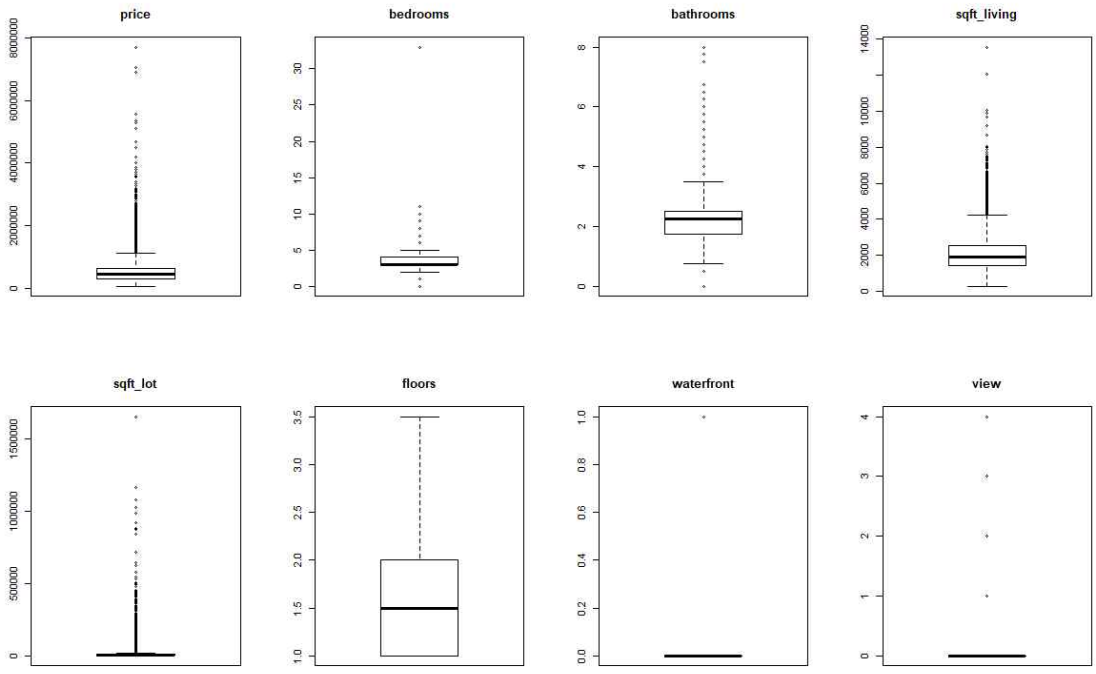
[Q1]

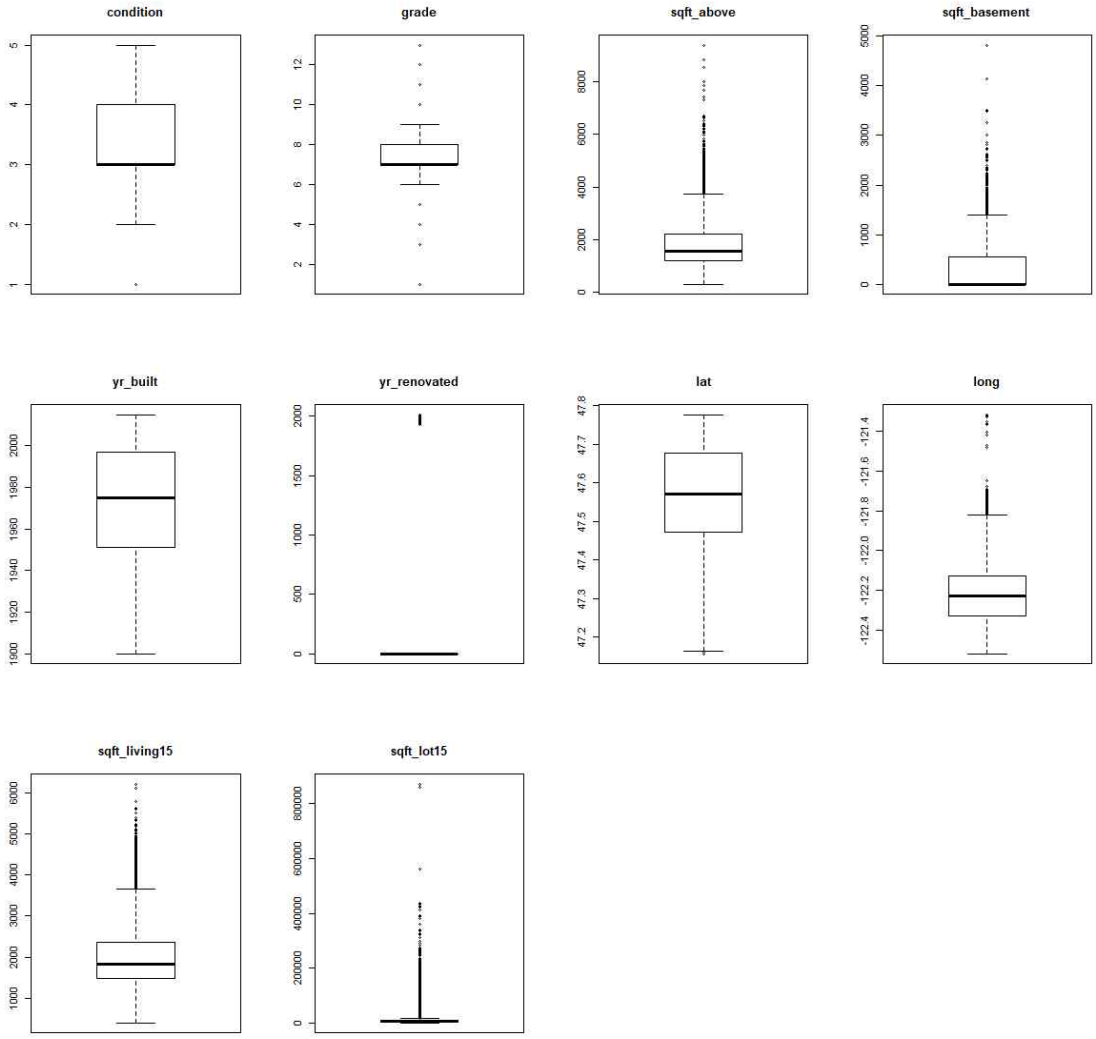
1번째, 2번째, 17번째 데이터인 id, date, zipcode는 유의미하지 않다. 이 값들은 유의미한 값이 아니라 임의로 지정된 숫자에 불과하므로 선형회귀의 입력 변수로서 적당하지 않다고 판단하여 입력변수에서 제거하였다.

[Q2]

	Mean	Standard Deviation	Skewness	Kurtosis
price	540088.141766529	367127.1964827	4.0237899	37.577262
bedrooms	3.370841623	0.9300618	1.9741625	52.052026
bathrooms	2.114757322	0.7701632	0.5110721	4.279329
sqft_living	2079.899736270	918.4408970	1.4714533	8.241603
sqft_lot	15106.967565817	41420.5115151	13.0591125	288.011596
floors	1.494308981	0.5399889	0.6161340	2.515112
waterfront	0.007541757	0.0865172	11.3843178	130.602691
view	0.234303428	0.7663176	3.3955139	13.890224
condition	3.409429510	0.6507430	1.0327330	3.525364
grade	7.656873178	1.1754588	0.7710497	4.190379
sqft_above	1788.390690788	828.0909777	1.4465641	6.401239
sqft_basement	291.509045482	442.5750427	1.5778555	5.714668
yr_built	1971.005135798	29.3734108	-0.4697728	2.342467
yr_renovated	84.402257900	401.6792400	4.5491776	21.696548
lat	47.560052519	0.1385637	-0.4852368	2.323566
long	-122.213896405	0.1408283	0.8849916	4.048981
sqft_living15	1986.552491556	685.3913043	1.1081044	4.596449
sqft_lot15	12768.455651691	27304.1796313	9.5060834	153.727957

1번에서 제거한 변수들을 제외한 변수들의 Mean, Standard Deviation, Skewness, Kurtosis는 위와 같다.





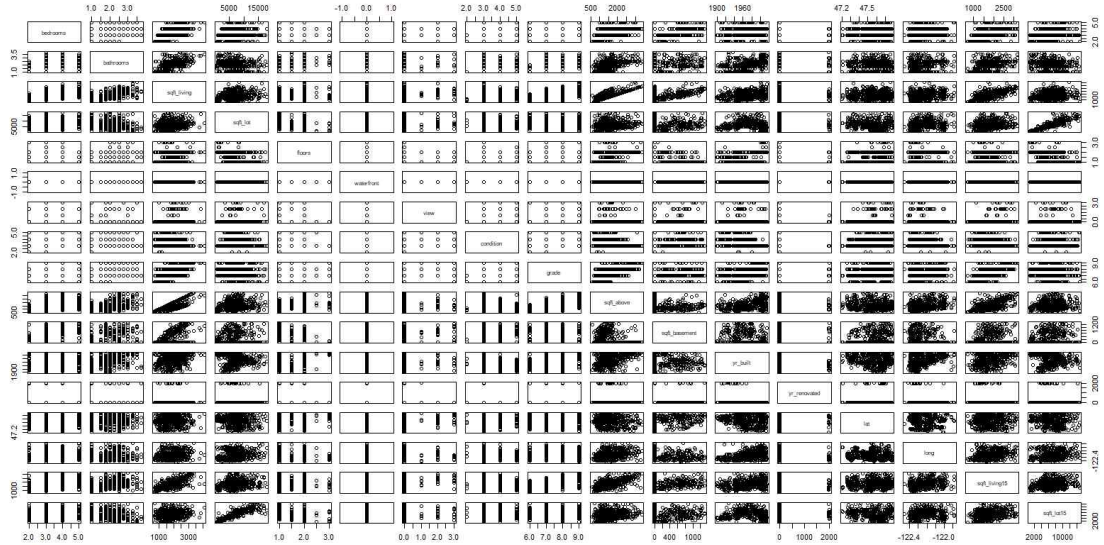
각 변수에 대한 skewness와 kurtosis, 그리고 box plot의 분포를 통해 정규분포를 파악해보자. skewness는 0에 가까워야 하고, kurtosis는 3에 가까워야 정규분포를 따른다고 할 수 있다. 다음과 같은 변수들이 정규분포를 따른다고 할 수 있다. bathrooms, floors, condition, grade, yr_built, lat, long, sqft_living15이 정규분포를 따른다고 할 수 있다.

[Q3]

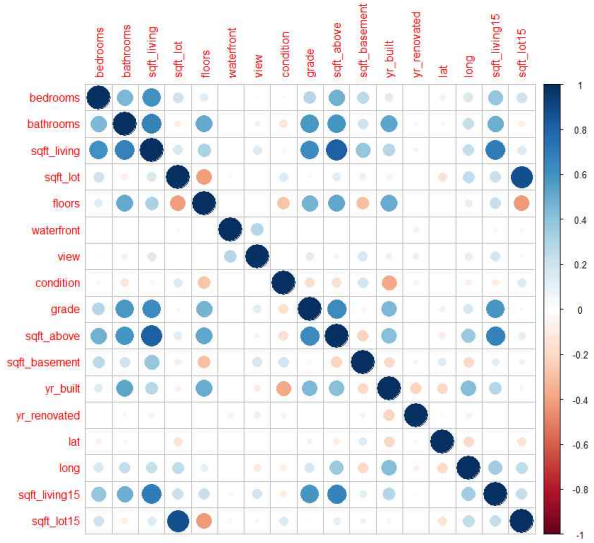
Q1에서 Q3 사이인 사분위간 범위(IQR)로 몸통을 구성하고 근접값들로 꼬리를 구성한다. 단위척도(step)는 1.5 x IQR이다. 안 울타리(inner fence)는 Q1에서 왼쪽으로 1 step만큼 간 것과 Q3에서 오른쪽으로 1 step만큼 간 것이다. 바깥 울타리(outer fence)는 Q1에서 왼쪽으로 2 step만큼 간 것과 Q3에서 오른쪽으로 2 step만큼 간 것이다. box plot을 기준으로 이상치는 whisker 범위보다 밖에 있으면 이상치로 정의하였다. 이에 따라 이상치를 제거하는 함수를 만들었고 이상치에 해당하는 객체들을 제거하였다. 단, waterfront, view, yr_renovated는 명목형 변수로 생각하고 제거하지 않았다. 이에 따라 원래 21613개의 데이터에서 15983개의 관측치를 가지는 데이터로 바뀌었다.

[Q4]

입력 변수들간의 scatterplot은 다음과 같다. 관측치의 개수가 너무 많아서 500개만 추출하여 scatterplot을 그렸다.



corrplot() 함수를 이용하여 correlation plot을 도시한 결과는 다음과 같다.

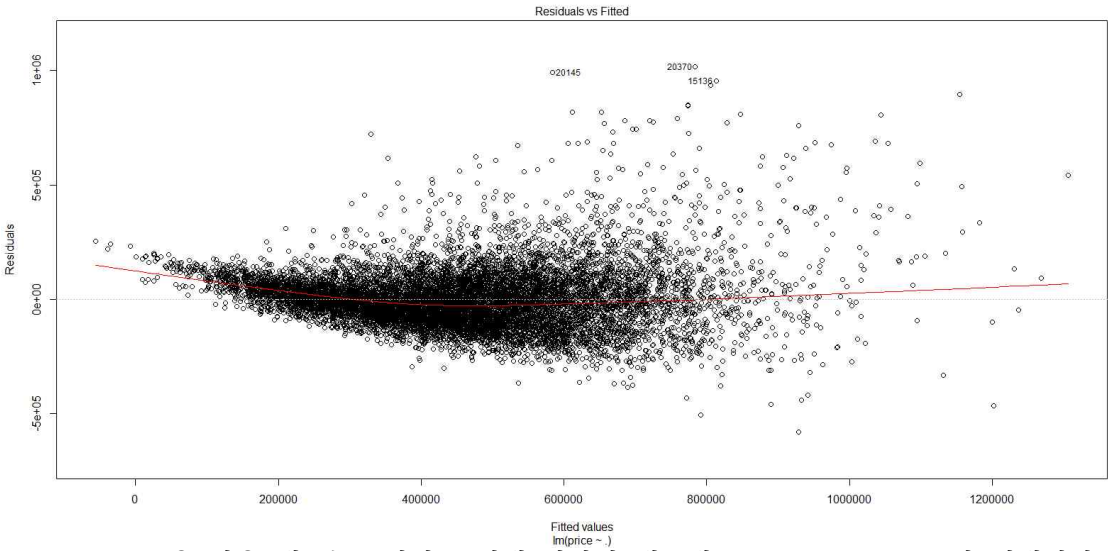


원의 색이 파랑색에 가까울수록 양의 상관관계를 갖는 것이고, 빨간색에 가까울수록 음의 상관관계를 갖는 것이다. 가장 큰 양의 상관관계를 갖는 두 입력변수 sqft_lot과 sqft_lot15 이다. 또한 sqft_living과 sqft_above도 높은 상관관계를 가진다. 이러한 상관관계를 가지는 변수들은 다중공선성을 유발할 수 있다.

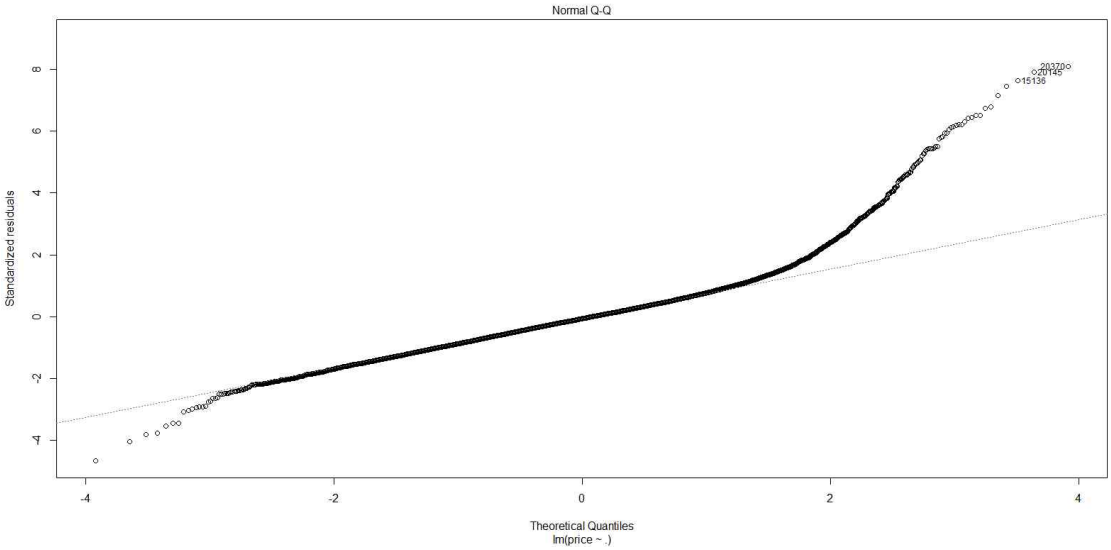
[Q5]

```
Residuals:
    Min      1Q  Median      3Q     Max
-582753  -76282   -7929   59246 1015946

Coefficients: (1 not defined because of singularities)
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.889e+07  1.644e+06 -11.489  < 2e-16 ***
bedrooms      -1.058e+04  2.000e+03  -5.290 1.24e-07 ***
bathrooms      2.018e+04  3.076e+03   6.559 5.66e-11 ***
sqft_living    9.477e+01  4.696e+00  20.180  < 2e-16 ***
sqft_lot      -2.642e+00  7.797e-01  -3.389 0.000705 ***
floors         6.886e+03  3.609e+03   1.908 0.056389 .
waterfront     3.535e+05  2.872e+04  12.310  < 2e-16 ***
view           3.701e+04  2.289e+03  16.166  < 2e-16 ***
condition      3.285e+04  2.054e+03  15.995  < 2e-16 ***
grade          8.781e+04  2.238e+03  39.237  < 2e-16 ***
sqft_above     1.343e+01  4.710e+00   2.852 0.004356 **
sqft_basement         NA         NA      NA       NA
yr_built      -2.058e+03  6.335e+01 -32.482  < 2e-16 ***
yr_renovated   2.445e+01  3.335e+00   7.332 2.42e-13 ***
lat            5.329e+05  9.054e+03  58.858  < 2e-16 ***
long           2.420e+04  1.265e+04   1.913 0.055754 .
sqft_living15  4.973e+01  3.777e+00  13.167  < 2e-16 ***
sqft_lot15    -6.160e+00  9.017e-01  -6.832 8.83e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 125500 on 11171 degrees of freedom
Multiple R-squared:  0.6511,    Adjusted R-squared:  0.6506
F-statistic:  1303 on 16 and 11171 DF,  p-value: < 2.2e-16
```

모든 입력변수를 사용하여 다중회귀모델을 사용하였을 때 위와 같은 결과를 얻는다. 이에 따른 Adj R-sq 값은 0.6506이다. 이 의미는 전체 데이터의 변동성 중에서 선형회귀모델이 약 65% 설명가능하다는 의미이다.



residual plot을 봤을 때 붉은 선이 수평에 가깝기 때문에 homoskedasticity가 위배되지 않았다고 할 수 있다.



qq plot에 따르면 residual 값들이 -2보다 작은 값에서부터 직선 위에 붙어있기 시작하면서 거의 1.7에 가까울 때 까지 직선 위에 붙어있다. 이를 통해 정규성 가정이 합당한 것을 알 수 있다.

[Q6]

```
Residuals:
    Min      1Q  Median      3Q     Max
-582753  -76282   -7929   59246 1015946

Coefficients: (1 not defined because of singularities)
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.889e+07  1.644e+06 -11.489  < 2e-16 ***
bedrooms      -1.058e+04  2.000e+03  -5.290 1.24e-07 ***
bathrooms      2.018e+04  3.076e+03   6.559 5.66e-11 ***
sqft_living    9.477e+01  4.696e+00  20.180  < 2e-16 ***
sqft_lot      -2.642e+00  7.797e-01  -3.389 0.000705 ***
floors         6.886e+03  3.609e+03   1.908 0.056389 .
waterfront     3.535e+05  2.872e+04  12.310  < 2e-16 ***
view           3.701e+04  2.289e+03  16.166  < 2e-16 ***
condition      3.285e+04  2.054e+03  15.995  < 2e-16 ***
grade          8.781e+04  2.238e+03  39.237  < 2e-16 ***
sqft_above     1.343e+01  4.710e+00   2.852 0.004356 **
sqft_basement         NA         NA      NA       NA
yr_built      -2.058e+03  6.335e+01 -32.482  < 2e-16 ***
yr_renovated   2.445e+01  3.335e+00   7.332 2.42e-13 ***
lat            5.329e+05  9.054e+03  58.858  < 2e-16 ***
long           2.420e+04  1.265e+04   1.913 0.055754 .
sqft_living15  4.973e+01  3.777e+00  13.167  < 2e-16 ***
sqft_lot15    -6.160e+00  9.017e-01  -6.832 8.83e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 125500 on 11171 degrees of freedom
Multiple R-squared:  0.6511,    Adjusted R-squared:  0.6506
F-statistic:  1303 on 16 and 11171 DF,  p-value: < 2.2e-16
```

유의수준 0.01에서 설명 변수들의 유의성을 판단했을 때 floors와 long 변수가 H0: beta = 0

을 reject하지 못한다. 또한 sqft_basement는 singularity에 의해 NA로 표현되었다. 이 의미는 변수들 사이에 다중공선성이 존재함을 알 수 있다. Price와 음의 상관관계를 갖는 변수로는 bedrooms, sqft_lot, yr_built, sqft_lot15가 있다. 이 외의 변수들은 양의 상관관계를 갖는 변수이다.

[Q7]

```
            RMSE      MAE     MAPE
House 125864.9 89185.05 20.7862
```

모든 변수를 사용하여 RMSE, MAE, MAPE를 구하였을 때 위와 같은 값을 얻는다. RMSE는 실제 price값과 모델이 예측한 price 값의 차이의 제곱을 루트 씌운 값의 평균이다. 이 값이 클수록 모델과 실제값 사이의 오차가 크다는 것이다. MAE는 단순히 실제값과 예측값 사이의 차이의 절대값의 평균이다. 이 의미는 약 89185$의 차이가 있다는 것이다. MAPE는 예측값과 실제값 사이의 차이가 약 20%정도 난다는 것이다.

[Q8]

lat(15), grade(10), yr_built(13), sqft_living(4), waterfront(7), view(8), condition(9)을 사용할 것이다. 이 변수들을 뽑은 이유는 t_value의 절대값이 높은 순서대로 뽑았고, 먼저 뽑힌 변수들과 높은 상관관계를 가지는 변수들은 t_value 값이 높더라도 뽑지 않았다.

[Q9]

```
Residuals:
    Min      1Q  Median      3Q     Max
-526515  -79385   -7652   64618 1033670

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.361e+07  4.675e+05  -50.51   <2e-16 ***
sqft_living  1.181e+02  2.609e+00   45.26   <2e-16 ***
waterfront   3.626e+05  2.956e+04   12.27   <2e-16 ***
view         4.093e+04  2.306e+03   17.75   <2e-16 ***
condition    2.412e+04  2.047e+03   11.79   <2e-16 ***
grade        1.056e+05  2.135e+03   49.46   <2e-16 ***
yr_built    -1.809e+03  5.110e+01  -35.41   <2e-16 ***
lat          5.582e+05  9.105e+03   61.31   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 129700 on 11180 degrees of freedom
Multiple R-squared:  0.6272,    Adjusted R-squared:  0.6269
F-statistic:  2687 on 7 and 11180 DF,  p-value: < 2.2e-16
```

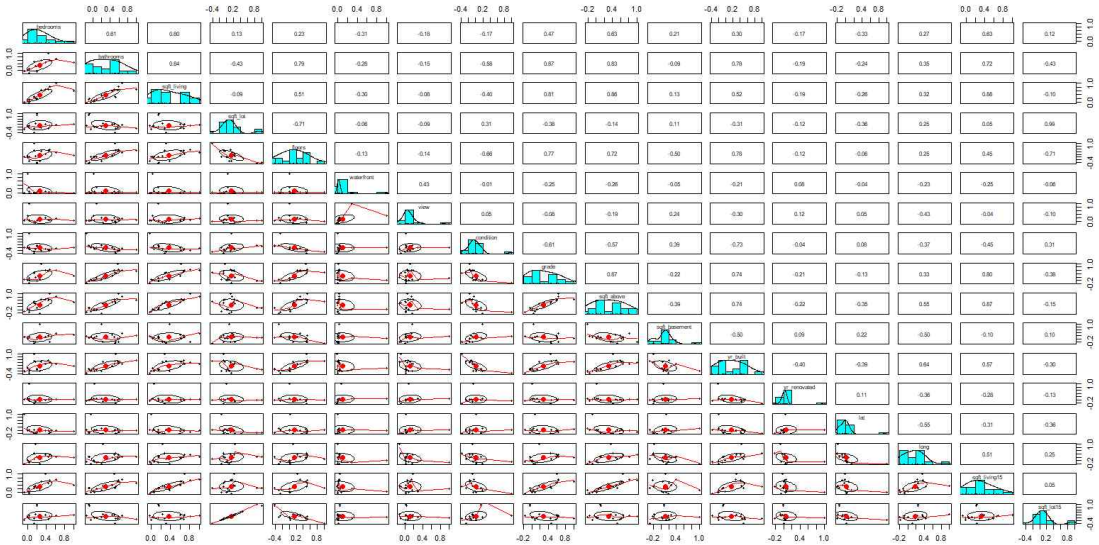
Adj R-sq 값은 0.6269로 변수를 많이 제거 하였지만 크게 낮아지지 않았고, 모든 변수의 P value가 0에 가깝다.

```
              RMSE      MAE     MAPE
House    125864.9 89185.05 20.78620
House_7  130425.9 94069.24 21.99657
```

House_7이 7개의 변수만으로 다시 세운 회귀의 오차 값들이다. 원래의 모델에서 많은 변수를 제거 했지만 세가지 regression evaluation criteria 값이 많이 변하지 않은것으로보아 변수를 제거한 모델이 꽤 좋은 모델임을 알 수 있다.

[Extra Question]

“psych” library를 사용하여 상관관계에 대한 다른 시각화를 시도해보았다.



upper triangular matrix는 각 변수간의 correlation 값을 나타내고 diagonal matrix는 각 변수의 정규성을 보여준다. 또한 lower triangular matrix는 각 변수의 관측치를 이차 평면에 도시하고 관계를 보여준 것이다. 양의 상관관계를 가지는 변수간에는 붉은 선이 우상향하는 모습을 보인다.