<다변량분석 리포트3>

2014170852 조영관

<1. MLR 구축을 위해 필요하지 않은 변수>

변수들을 살펴보았습니다.

ID와 DATE 그리고 ZIP, LAT, LON, sqft_living, sqft_lot은 불필요하다고 생각하여 제거하였습니다.

ID와 DATE의 경우 각각 ID와 시기를 표현하는 일련의 코드이므로 회귀분석을 함에 있어 불필요합니다. 그리고 ZIP의 주소번호 역시 불필요합니다. LAT, LON은 위도와 경도인데, 위치가 어디 있는지에 대한 지표는 집의 가격을 형성하는 것에 불필요하다 생각하여 제거하였습니다.

Sqft_living, sqft_lot을 제거한 이유는, 이미 sqft_living15, sqft_lot15 즉 최근 측정 데이터가 이미 존재하므로 불필요하다고 생각하여 제거하였습니다.

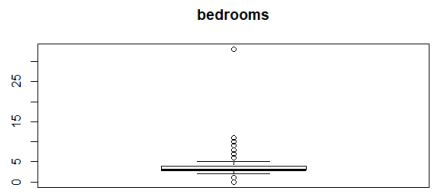
<2. 단변량 통계량 계산 및 BOX PLOT 도시, 정규분포를 따르는 변수와 그 근거>

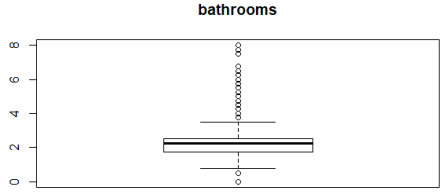
	bedrooms	bathroom	floors	condition	grade	sqft_above	sqft_baser	yr_built	sqft_living	sqft_lot15
mean	3.37	2.11	1.49	3.41	7.66	1788.39	291.51	1971.01	1986.55	12768.46
standard deviation	0.93	0.77	0.54	0.65	1.18	828.09	442.58	29.37	685.39	27304.18
skewness	1.97	0.51	0.62	1.03	0.77	1.45	1.58	-0.47	1.11	9.51
kurtosis	52.05	4.28	2.52	3.53	4.19	6.4	5.72	2.34	4.6	153.73

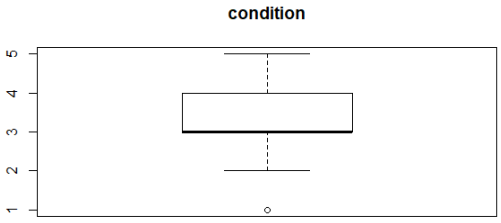
개별 입력변수들의 통계량은 다음과 같습니다.

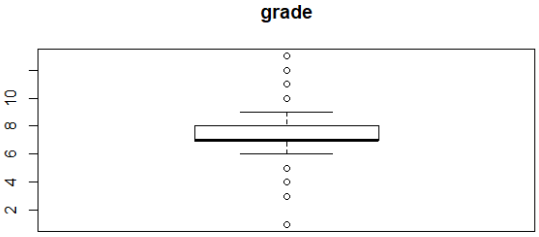
범주형 변수는 제외하고 산출하였습니다.

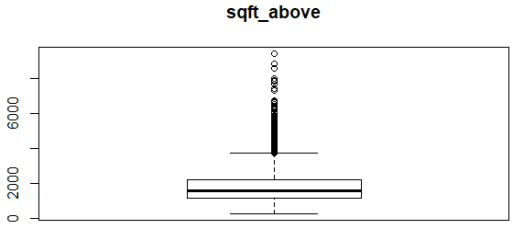
Box plot을 도시해보겠습니다. 하나로 합치니 단위 차이가 너무 많이 나서 box plot을 제대로 보기가 어려워서 따로따로 분리하였습니다.

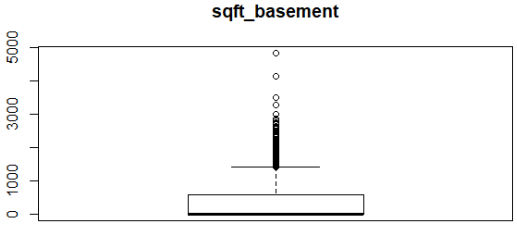


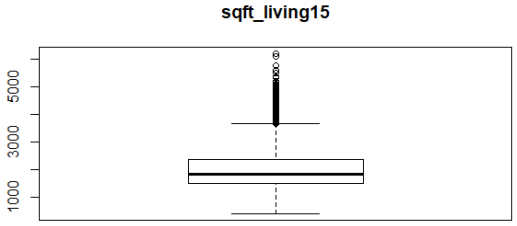


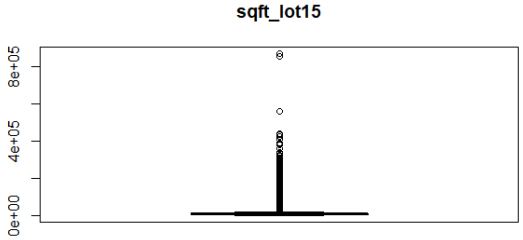

condition
1
2
3
4
5

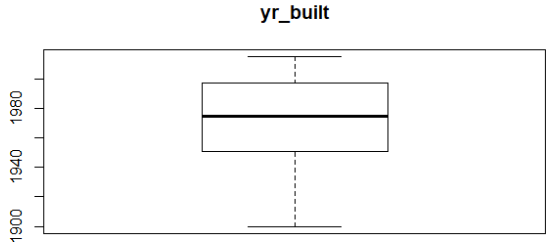

grade
2
4
6
8
10

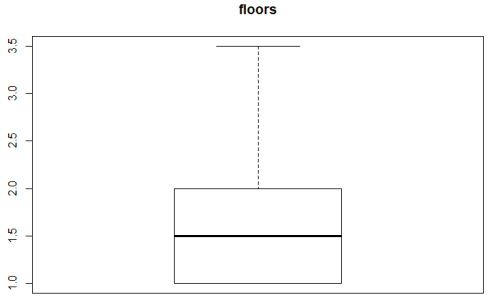

sqft_above
0
2000
6000


sqft_basement
0
1000
3000
5000

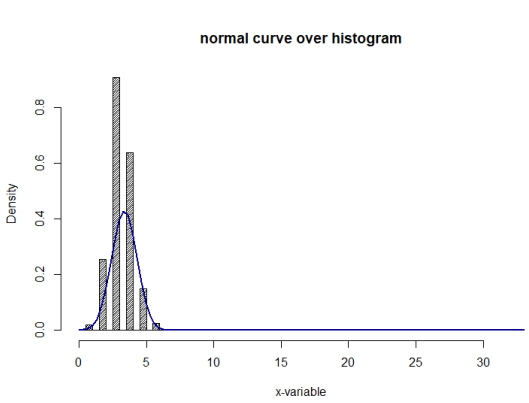

sqft_living15
1000
3000
5000

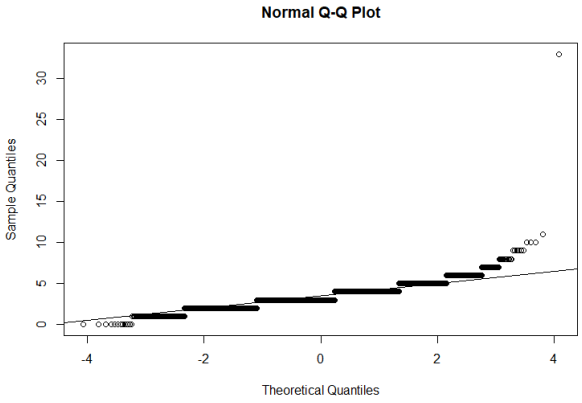

sqft_lot15
0e+00
4e+05
8e+05


yr_built
1900
1940
1980

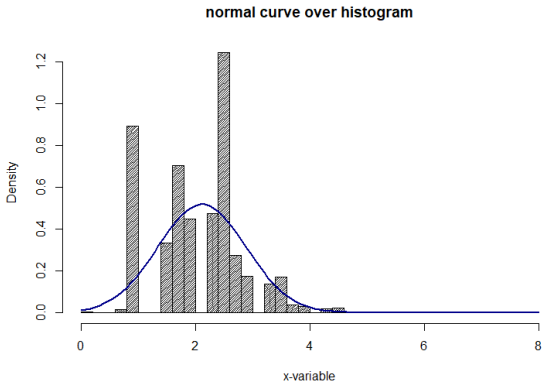

floors
1.0
1.5
2.0
2.5
3.0
3.5

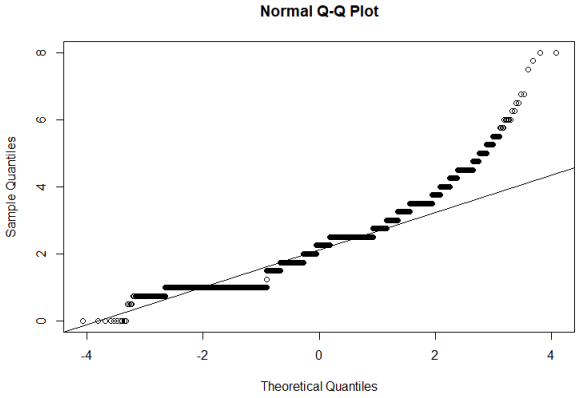
정규분포를 따르는 변수가 얼마나 있는지 확인해보겠습니다.



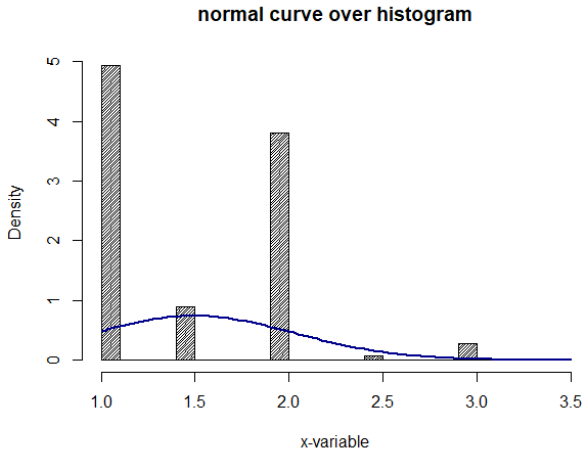


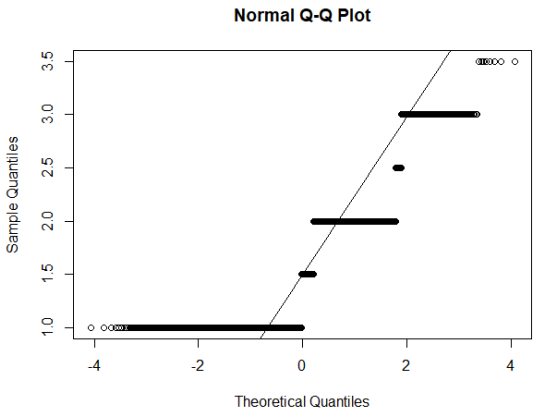
먼저 bedroom의 경우 히스토그램과 Q-Q PLOT을 보면 어느 정도 정규분포를 따르고 있다고 할 수 있습니다.



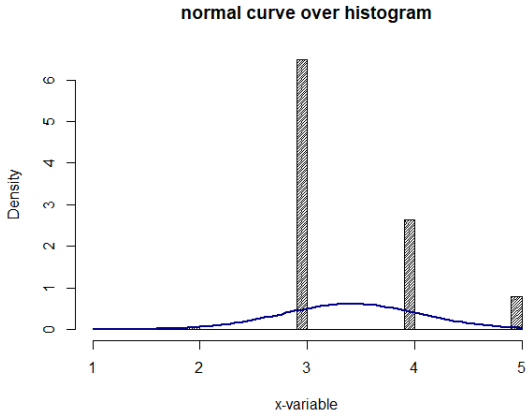


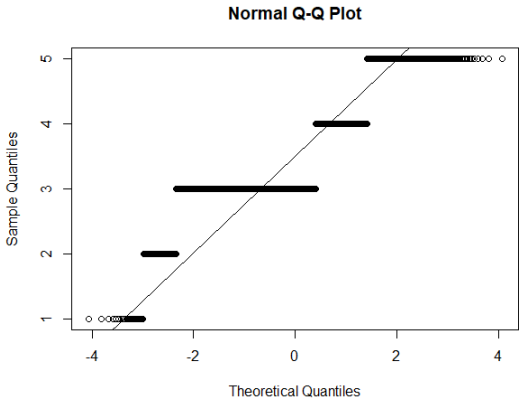
Bathroom의 경우 histogram과 q-q plot을 봤을 때 비슷하지는 않더라도, 벨 모양 형태를 보이며, qqline 위에 점들이 대체로 잘 배열되어 있으므로 정규성을 따르고 있다고 볼 수 있습니다.



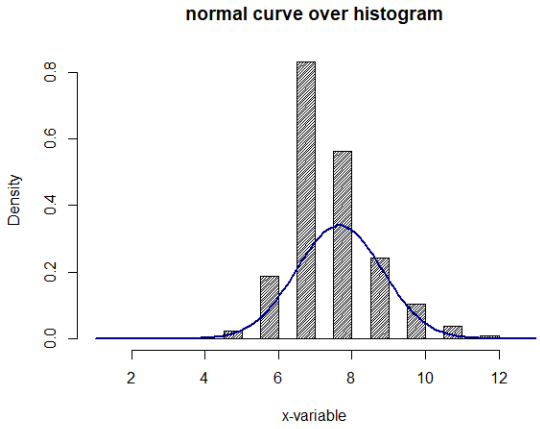


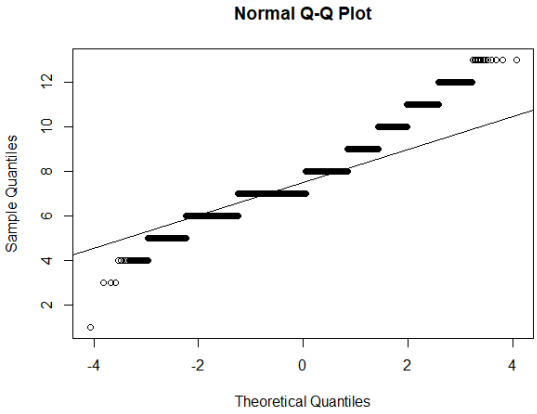
Floors 변수의 경우 정규분포를 따르지 않음을 확인할 수 있습니다.



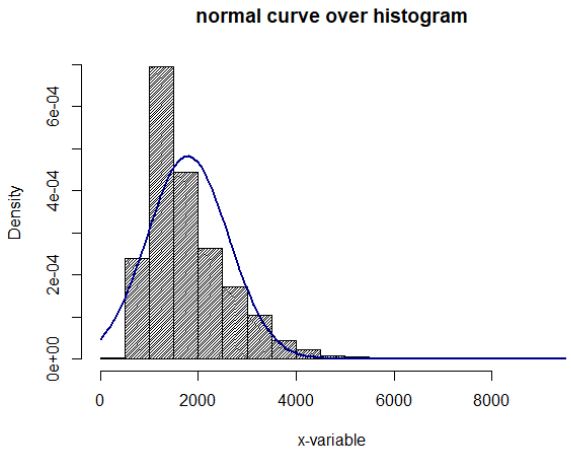


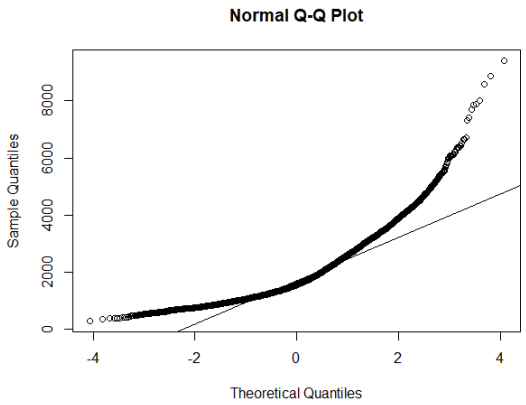
Condition 변수의 경우 정규분포를 따르지 않습니다.





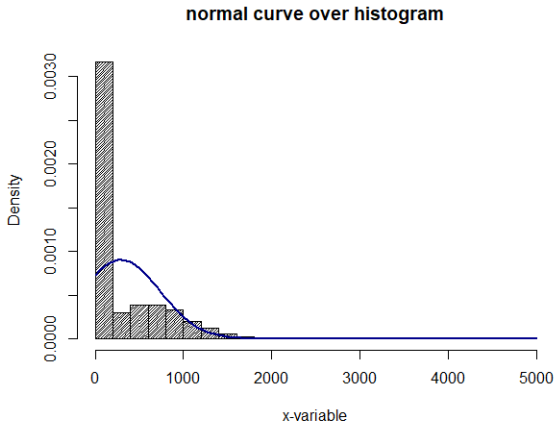
Grade 변수의 경우 qq plot을 보면 정규분포를 따르고 있다고 하기에 조금 rough 하지만 히스토그램이 벨 모양 형태를 띄고 있고, skewness와 kurtosis가 정규분포와 비슷하기에 정규성을 가진다고 할 수 있습니다.

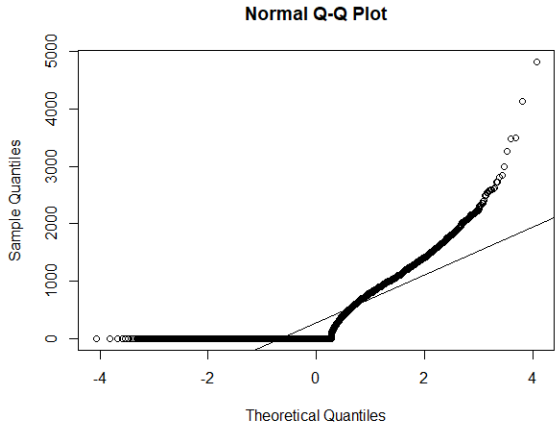




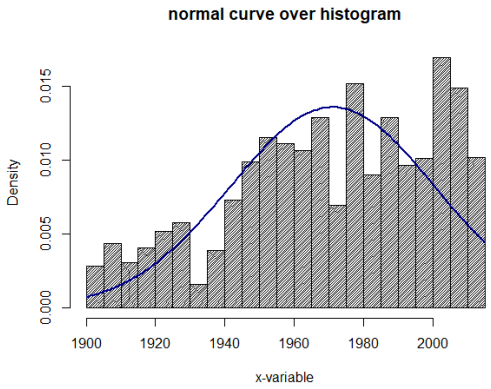
Sqft_above 변수는 qqplot을 보면 점 배열이 정확한 직선은 아니고 약간 휘어져 있지만, 대체로

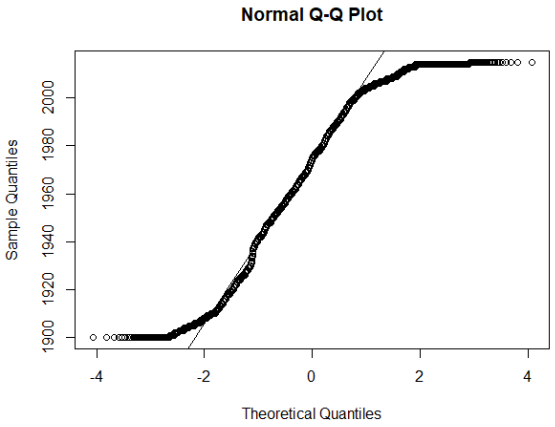
점 배열들이 선 위에 있으므로 정규성을 가진다고 할 수 있습니다.



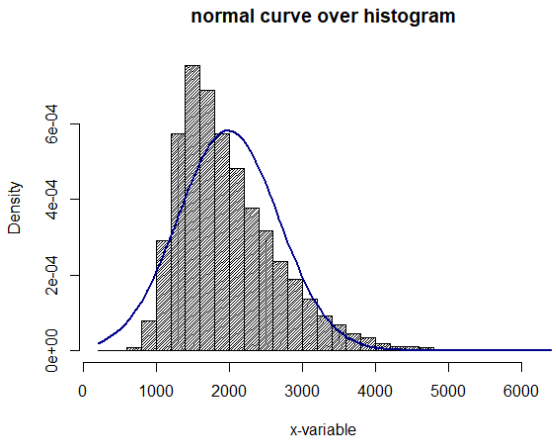


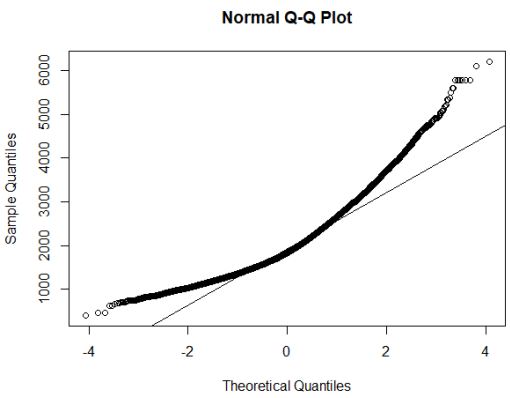
Sqft_basement 변수는 정규분포를 따르지 않습니다.





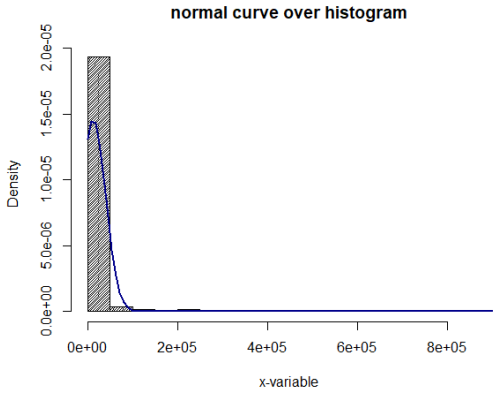
Yr_built 변수는 히스토그램이 벨 모양에 근사하지는 않지만, qq plot을 보면 선 위에 점들이 잘 배열되어 있으며, skewness, kurtosis 값이 정규분포에 근사하므로 정규분포를 따릅니다.

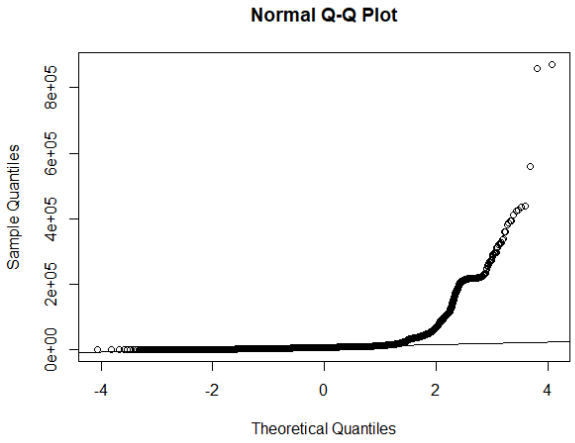




Sqft_living15 변수는 histogram의 모양, qqplot의 점 배열을 봤을 때, 정확히는 아니더라도 대략적

으로 정규분포에 근사함을 알 수 있습니다.





Sqft_lot15 변수는 정규분포를 따르지 않습니다.

결론적으로, bedroom, bathroom, grade, sqft_above, yr_built, sqft_living15가 정규분포에 근사함을 확인할 수 있었습니다.

그 이외의 변수들은 histogram이 정규분포의 벨 모양을 띄지 않거나, q-q plot을 도시해본 결과 정규분포 선 위에 제대로 정렬되어 있지 않음을 확인할 수 있습니다.

<3. 각 변수들에 대한 이상치 조건 정의, 해당 객체 데이터 제거>

Box plot을 살펴보면, box로 형성되어 있는 영역에서 벗어난 점들이 몇 개 보입니다.

이 점들은 분포에서 벗어난 이상치로 분류하여야 합니다.

그래서 box plot 기준으로 이상치로 분류되는 데이터들을 제거하였습니다.

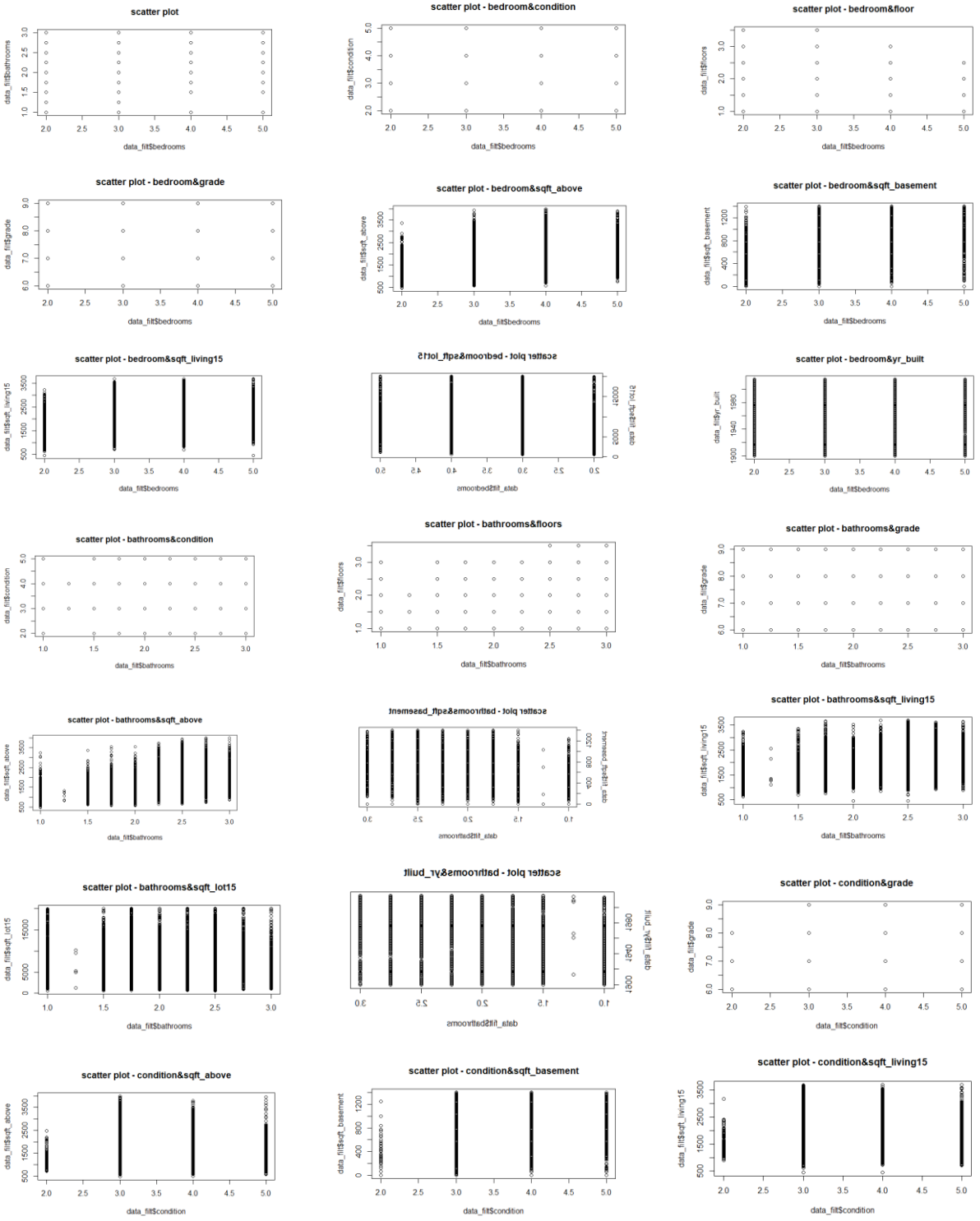
컴퓨터에 내장된 함수가 걸러내는 것이기 때문에, 그 정확한 기준은 알 수 없지만,

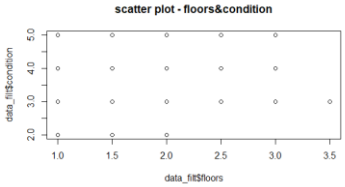
대략적인 이상치 영역들을 적어보았습니다.

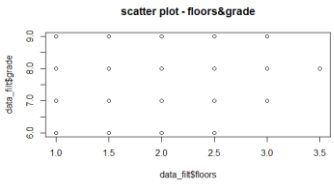
```
#----------------------이상치 파악하기
# bedrooms -> 2~5
# bathrooms -> 1~3
# floors
# view
# condition -> 2~5
# grade -> 6~9
# sqft_above ->  400~ 4000
# sqft_basement -> 0 ~ 1400
# yr_built
# sqft_living15 -> 300 ~ 3700
# sqft_lot15 -> 0 ~ 20000
```

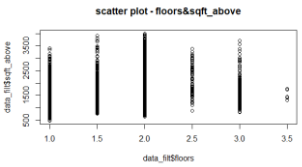
대략 이 범주 안에서 벗어나면 이상치로 분류됩니다.

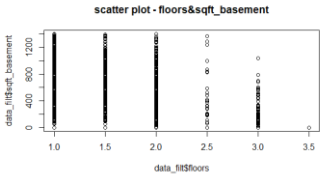
<4. Scatter plot 및 Correlation plot 그리기, 강한 상관관계를 가지고 있는 변수 파악>

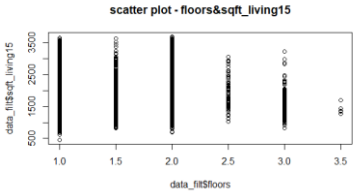


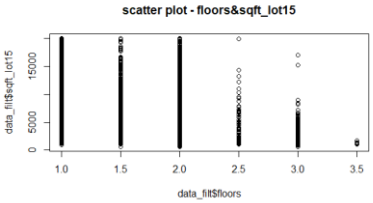

scatter plot - floors&condition
data_filt$condition
data_filt$floors

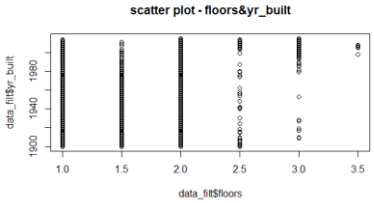

scatter plot - floors&grade
data_filt$grade
data_filt$floors

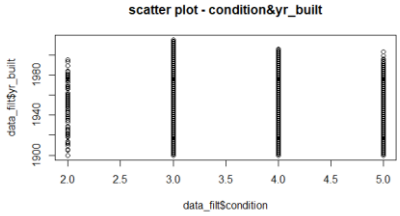

scatter plot - floors&sqft_above
data_filt$sqft_above
data_filt$floors

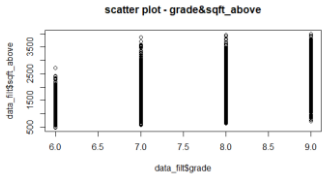

scatter plot - floors&sqft_basement
data_filt$sqft_basement
data_filt$floors

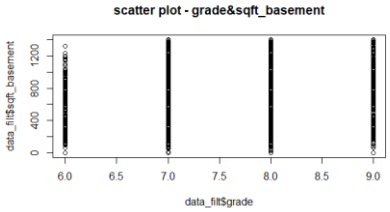

scatter plot - floors&sqft_living15
data_filt$sqft_living15
data_filt$floors

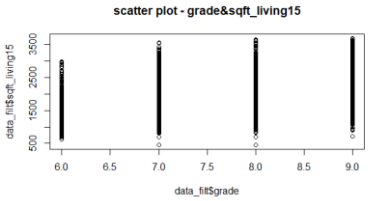

scatter plot - floors&sqft_lot15
data_filt$sqft_lot15
data_filt$floors

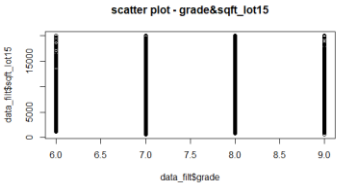

scatter plot - floors&yr_built
data_filt$yr_built
data_filt$floors

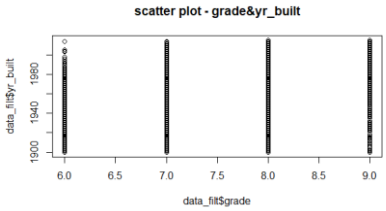

scatter plot - condition&yr_built
data_filt$yr_built
data_filt$condition

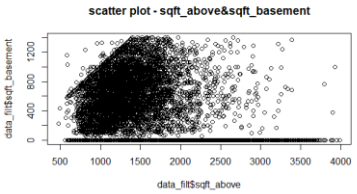

scatter plot - grade&sqft_above
data_filt$sqft_above
data_filt$grade

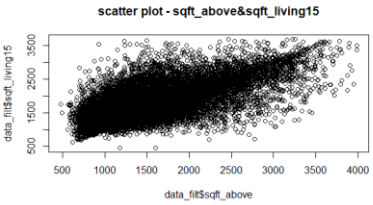

scatter plot - grade&sqft_basement
data_filt$sqft_basement
data_filt$grade

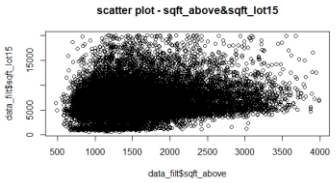

scatter plot - grade&sqft_living15
data_filt$sqft_living15
data_filt$grade

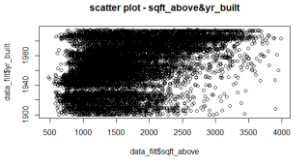

scatter plot - grade&sqft_lot15
data_filt$sqft_lot15
data_filt$grade

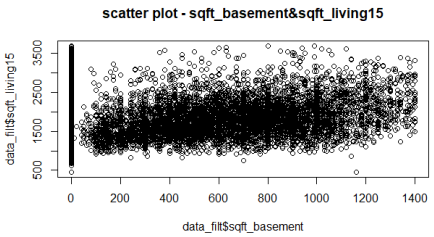

scatter plot - grade&yr_built
data_filt$yr_built
data_filt$grade

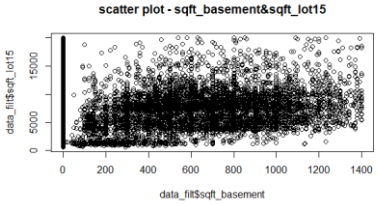

scatter plot - sqft_above&sqft_basement
data_filt$sqft_basement
data_filt$sqft_above

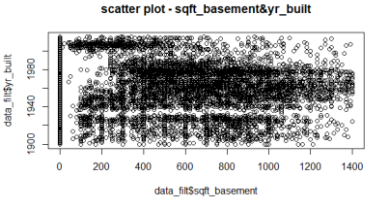

scatter plot - sqft_above&sqft_living15
data_filt$sqft_living15
data_filt$sqft_above


scatter plot - sqft_above&sqft_lot15
data_filt$sqft_lot15
data_filt$sqft_above

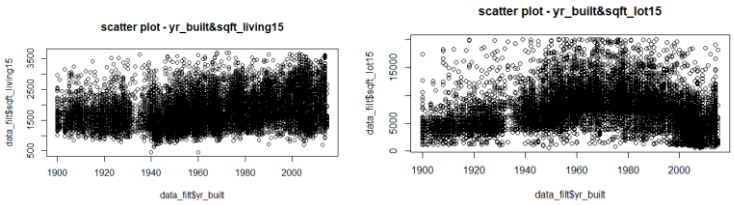

scatter plot - sqft_above&yr_built
data_filt$yr_built
data_filt$sqft_above


scatter plot - sqft_basement&sqft_living15
data_filt$sqft_living15
data_filt$sqft_basement


scatter plot - sqft_basement&sqft_lot15
data_filt$sqft_lot15
data_filt$sqft_basement

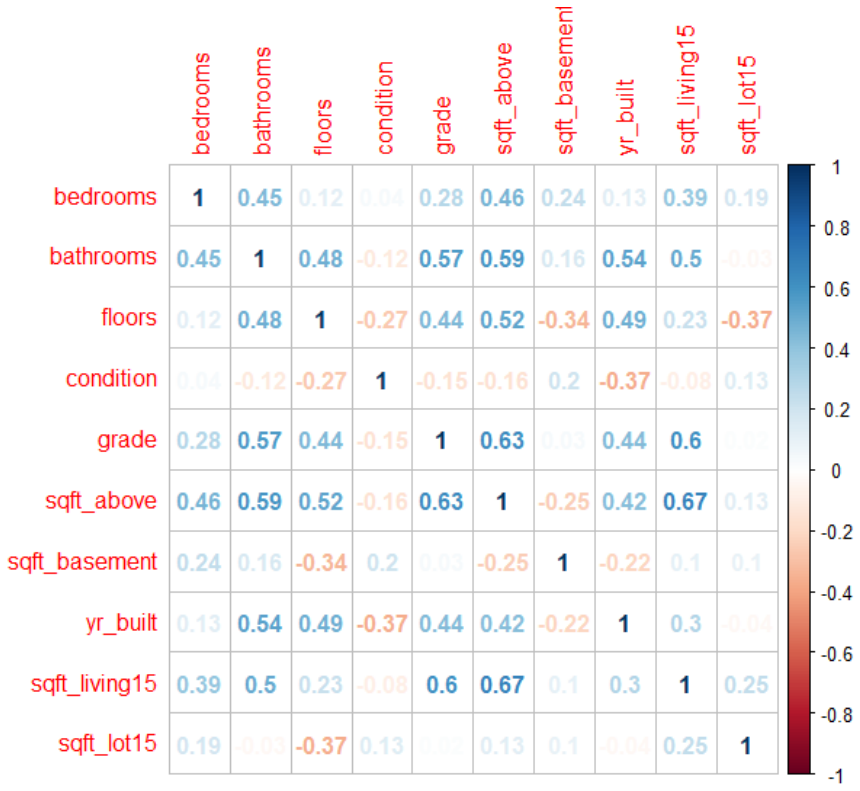

scatter plot - sqft_basement&yr_built
data_filt$yr_built
data_filt$sqft_basement


scatter plot - sqft_living15&sqft_lot15
data_filt$sqft_lot15
data_filt$sqft_living15



Scatter plot은 위와 같습니다. 한 화면에 담으니 개별 scatter plot이 너무 작아져서 잘 보이지 않아 따로 분리하여 나열하였습니다.

Bedroom, bathroom과 같은 정수형 변수가 있다 보니 scatter plot이 직선의 형태로 보이는 그래프가 꽤 있습니다. 다음으로 corr plot을 살펴봅시다.



Corr plot을 통해 입력변수 간 correlation을 파악할 수 있습니다.

이 중 강한 상관관계를 가지는 변수들을 찾아봅시다.

Bathrooms 변수의 경우 grade, sqft_above, yr_built와 높은 편의 상관관계를 가집니다.

Floors의 경우 sqft_above의 변수와 높은 편의 상관관계를 가집니다.

Grade 변수의 경우 sqft_above,, sqft_living15와 높은 편의 상관관계를 가집니다.

Sqft_above 변수의 경우 sqft_living15와 높은 편의 상관관계를 가집니다.

그 이외의 변수들 관계는 높지 않습니다.

<5. training data, test data로 분할한 후 MLR 모델 학습하기. 모델 선형성 판단, residual plot, Q-Q PLOT 도시하기>

```
Residuals:
    Min      1Q  Median      3Q     Max
-542223  -87663   -7031   74216 1226814

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.963e+06  1.305e+05  38.042  < 2e-16 ***
bedrooms      -2.109e+04  2.252e+03  -9.364  < 2e-16 ***
bathrooms      1.832e+04  3.559e+03   5.147 2.69e-07 ***
floors         3.562e+04  3.980e+03   8.950  < 2e-16 ***
condition      2.315e+04  2.292e+03  10.098  < 2e-16 ***
grade          1.039e+05  2.469e+03  42.089  < 2e-16 ***
sqft_above     8.817e+01  4.570e+00  19.293  < 2e-16 ***
sqft_basement  1.225e+02  5.259e+00  23.288  < 2e-16 ***
yr_built      -2.840e+03  6.742e+01 -42.122  < 2e-16 ***
sqft_living15  5.569e+01  4.161e+00  13.382  < 2e-16 ***
sqft_lot15    -7.922e+00  5.354e-01 -14.797  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 139000 on 10832 degrees of freedom
Multiple R-squared:  0.5019,    Adjusted R-squared:  0.5015
F-statistic:  1092 on 10 and 10832 DF,  p-value: < 2.2e-16
```

Training data와 test data를 7대 3으로 분할한 후 MLR 모델을 학습하였습니다.

위와 같은 결과를 얻을 수 있었습니다.

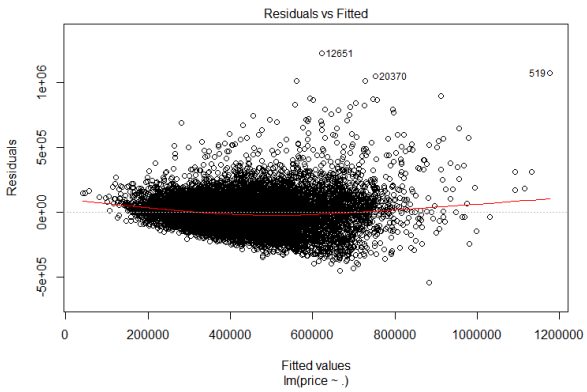
모든 변수들이 P-VALUE가 충분히 작아 유의한 변수임을 확인할 수 있습니다.

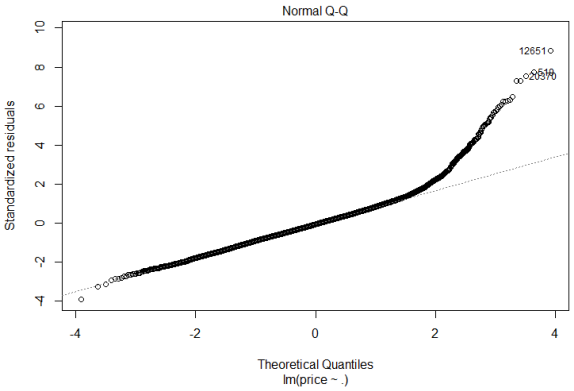
Adjusted R-SQUARED 값은 0.5015임을 알 수 있습니다.

즉, 위 모델이 실제 데이터의 변동성에 50.15% 정도의 설명력을 가지고 있다고 말할 수 있습니다.

즉, 위 데이터가 선형성을 가지고 있다고 보기에는 수치가 작습니다.

잔차 PLOT과 Q-Q PLOT을 확인해보겠습니다.





RESIDUAL PLOT을 보면 비교적 random하게 분포되어 있다고 볼 수도 있어서 등분산성이 있다고 할 수도 있지만, 개인적인 정성적 평가로는 왼쪽은 세로 두께가 짧은데 반해, 오른쪽으로 갈수록 그 길이가 길어집니다. 즉 데이터가 선형성을 띄지 않는다고 볼 수 있습니다. (이에 관해서는 아래 extra 분석에서 보강해보았습니다.)

Q-Q PLOT을 보면 위 모델 데이터는 어느 정도 정규성을 가짐을 확인할 수 있습니다.

따라서 정규성은 어느 정도 만족하지만, RESIDUAL PLOT을 통해 등분산성이 있다고 볼 수 없으므로 ORDINARY LEAST SQUARE SOLUTION을 사용하기에 적합하지 않다고 결론 내릴 수 있습니다.

<6. 유의수준 0.01에서 유의미한 변수들 확인, 해당 변수의 PRICE와의 상관관계>

위의 결과를 통해 알 수 있듯이, 모든 변수들이 P-VALUE가 충분히 작아 유의수준 0.01에서 귀무가설이 기각됨을 알 수 있습니다. 즉, 다시 말해 모든 변수들은 TARGET VARIABLE인 PRICE에 유의미한 영향을 미칩니다. 각각의 변수들의 양, 음 관계를 살펴보겠습니다.

Bedrooms, yr_built, sqft_lot15 이렇게 3개의 변수는 음의 관계를 가집니다.

그 이외 나머지 변수들은 양의 상관관계를 가짐을 알 수 있습니다.

<7. test data set으로 MAE, MAPE, RMSE 계산하기>

	RMSE	MAE	MAPE
House price	140598.9	103576.9	25.81793

결과를 산출해보면 다음과 같이 나옵니다.

즉, 테스트 데이터에 관해 모델 성능을 확인했을 때, RMSE는 140598.9 만큼의 예측 값과 실제 데이터 간 ERROR를 가지며 MAE는 103576.9 만큼의 예측 값과 실제 데이터간 ERROR를 가지고, 마지막으로 MAPE는 예측값과 실제 데이터 간 차이 ERROR의 비율이 25.81793% 정도임을 알 수

있습니다.

<8. 7개의 입력변수만을 사용하여 모델 구축 -> 어떤 변수를 선택할까?>

판단 근거

첫 번째로, 유의수준 0.01에서 변수 P-VALUE가 귀무 가설을 기각하는지를 살펴봅니다.

위에서 확인한 대로, 모든 변수들이 P-VALUE 값이 충분히 작으므로 제거할 변수는 없습니다.

두 번째로, 변수들 간에 강한 상관관계가 있으면 모델의 성능을 저하시킵니다.

Multicollinearity 문제가 발생하기 때문입니다.

따라서, 강한 상관관계에 있는 변수들을 살펴보고, 그 중 덜 중요한 비중을 차지하는 변수를 제거하였습니다. 그래서 그 결과 bathrooms, grade, sqft_above 변수를 제거한 후 7개의 변수를 남겼습니다.

<9. 선택한 7개의 변수를 활용해 MLR 모델을 학습하고 결과 살펴보기>

```
Call:
lm(formula = price ~ ., data = house_trn_data7)

Residuals:
    Min      1Q  Median      3Q     Max
-517602 -105449  -15815   83401 1518583

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.742e+06  1.311e+05  28.551   <2e-16 ***
bedrooms       5.152e+03  2.279e+03   2.261   0.0238 *
floors         1.148e+05  3.995e+03  28.745   <2e-16 ***
condition      2.999e+04  2.584e+03  11.608   <2e-16 ***
sqft_basement  1.129e+02  4.798e+00  23.519   <2e-16 ***
yr_built      -1.978e+03  6.690e+01 -29.572   <2e-16 ***
sqft_living15  1.871e+02  3.557e+00  52.604   <2e-16 ***
sqft_lot15    -7.084e+00  5.984e-01 -11.838   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 157000 on 10835 degrees of freedom
Multiple R-squared:  0.3672,    Adjusted R-squared:  0.3668
F-statistic: 898.4 on 7 and 10835 DF,  p-value: < 2.2e-16
```

MLR 모델을 학습한 결과는 위와 같았습니다.

TARGET VARIABLE에 유의미한 영향을 미치는 변수를 제거하니, Adjusted R-SQUARE가 낮아졌습니다.

	RMSE	MAE	MAPE
House price_7	162123.3	118461.1	29.44765

RMSE, MAE, MAPE를 산출한 결과는 다음과 같습니다.

즉, 위의 7개의 변수로 모델을 학습했을 때, 예측 값과 실제 데이터의 차이 ERROR는 RMSE로 162123.3 만큼, MAE는 118461.1만큼, 그리고 MAPE는 29.44765%만큼의 ERROR가 발생함을 알 수 있습니다.

TARGET VARIABLE에 유의미한 영향을 미치는 변수를 제거한 것이, 강한 상관관계를 가지는 변수를 제거하는 것보다 모델의 성능을 더 저하시키는 결과를 가져왔다고 결론 내릴 수 있습니다.

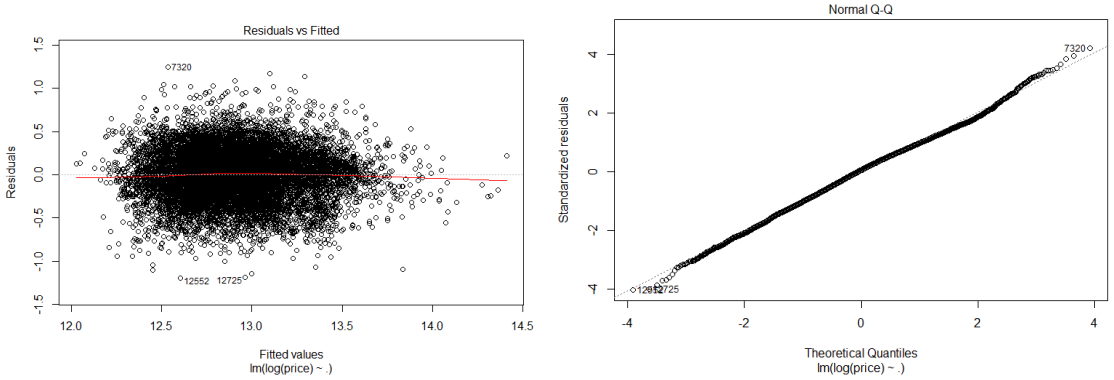
<extra 분석>

Log로 y값 변형하기

위 분석을 보면 알 수 있듯이, 정규성은 어느 정도 만족하지만 등분산성에 있어 잔차 plot을 보면 문제가 있습니다.

따라서, target variable인 y값에 변형을 주겠습니다.

저는 log를 씌워 분석을 진행하였습니다.



위와 같은 결과를 얻을 수 있습니다.

Residual plot을 보면, 잔차들이 규칙없이 random하게 분포되어 있음을 확인할 수 있습니다. 그리고 오른쪽 q-q plot을 확인해보면 점들이 정규 qqline에 잘 분포하고 있음을 확인할 수 있습니다.

따라서, 정규성과 등분산성을 만족합니다. 따라서, 일반 선형이 아닌 log를 씌워 결과를 도출하는 것이 더 타당합니다.

위 방법을 이용해 도출한 결과를 확인해봅시다.

```
Call:
lm(formula = log(price) ~ ., data = house_trn_data7)

Residuals:
     Min       1Q   Median       3Q      Max
-1.20011 -0.24182  0.01064  0.22907  1.43921

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.915e+01  2.804e-01  68.283   <2e-16 ***
bedrooms       9.393e-03  4.875e-03   1.927    0.054 .
floors         2.475e-01  8.545e-03  28.960   <2e-16 ***
condition      6.141e-02  5.527e-03  11.113   <2e-16 ***
sqft_basement  2.612e-04  1.026e-05  25.445   <2e-16 ***
yr_built      -3.800e-03  1.431e-04 -26.556   <2e-16 ***
sqft_living15  4.045e-04  7.608e-06  53.162   <2e-16 ***
sqft_lot15    -1.917e-05  1.280e-06 -14.975   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3359 on 10835 degrees of freedom
Multiple R-squared:  0.3775,    Adjusted R-squared:  0.3771
F-statistic: 938.5 on 7 and 10835 DF,  p-value: < 2.2e-16
```

결과는 위와 같습니다. 위에서 했던 7개의 변수를 가지고 진행한 결과와 비교해보면,

조금이지만 y가 선형일때에 비해 Adjusted R-SQUARE가 상승했음을 확인할 수 있습니다.

변수를 선택하는 방법

Forward, backward, stepwise selection 세 가지가 있습니다.

Forward는 하나씩 변수를 추가하는 것 (제일 중요한 것부터)

Backward는 하나씩 변수를 제거하는 것

Stepwise는 forward, backward 둘 다 이용하는 것입니다.

일단 위의 데이터의 경우, 결론적으로 선택된 모든 변수들이 target variable에 유의미한 영향을 미친다는 것이 확인되었으므로, 사실 상 위의 방법을 통해 결과물을 내도, 모든 변수들이 선택될 것입니다. 다만, 각 방법이 어떤 순으로 변수를 선택하는지, 빼는지를 확인해볼 수 있습니다.

```
Step:  AIC=372792.4
price ~ grade
```

```
Step:  AIC=370214.1
price ~ grade + yr_built
```

```
Step:  AIC=369113.3
price ~ grade + yr_built + bathrooms
```

```
Step:  AIC=368550.6
price ~ grade + yr_built + bathrooms + sqft_living15
```

```
Step:  AIC=368042.3
price ~ grade + yr_built + bathrooms + sqft_living15 + sqft_lot15
```

```
Step:  AIC=367864.7
price ~ grade + yr_built + bathrooms + sqft_living15 + sqft_lot15 +
    sqft_basement
```

```
Step:  AIC=367312.7
price ~ grade + yr_built + bathrooms + sqft_living15 + sqft_lot15 +
    sqft_basement + sqft_above
```

```
Step:  AIC=367183.9
price ~ grade + yr_built + bathrooms + sqft_living15 + sqft_lot15 +
    sqft_basement + sqft_above + condition
```

```
Step:  AIC=367083.5
price ~ grade + yr_built + bathrooms + sqft_living15 + sqft_lot15 +
    sqft_basement + sqft_above + condition + floors

Step:  AIC=366985.9
price ~ grade + yr_built + bathrooms + sqft_living15 + sqft_lot15 +
    sqft_basement + sqft_above + condition + floors + bedrooms
```

위와 같은 순서로 변수를 선택함을 알 수 있습니다.