Multivariate Data Analysis

Assignment #3

MLR: House price data

과목명	다변량분석
담당교수	강필성 교수님
제출일	2019-04-25
이름	박지원
학과명	산업경영공학부
학번	2014170856

목차

Introduction

Dataset: House Sales in King County, USA, kc_house_data.csv

해당 데이터셋은 미국 King Coynty에서 2014년 5월부터 2015년 5월까지 거래된 주택들에 대한 정보 및 가격이 포함되어 있습니다. 각 변수에 대한 설명은 제공된 URL에서의 Column 항목을 통해 확인할 수 있습니다. 이 중 세 번째 항목인 price가 MLR 모형의 target variable입니다.

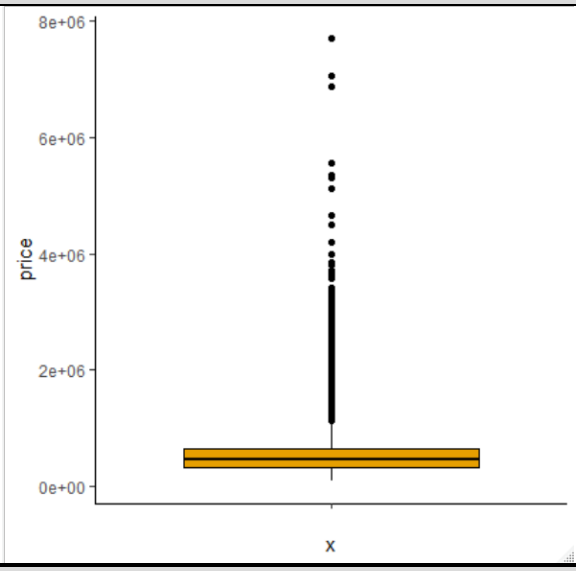
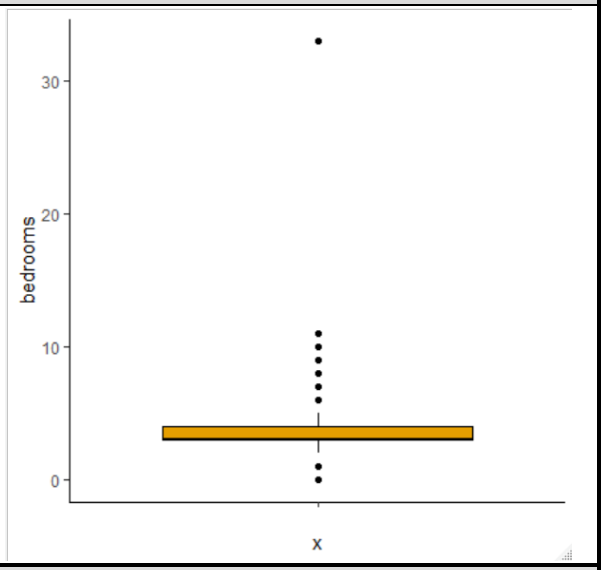
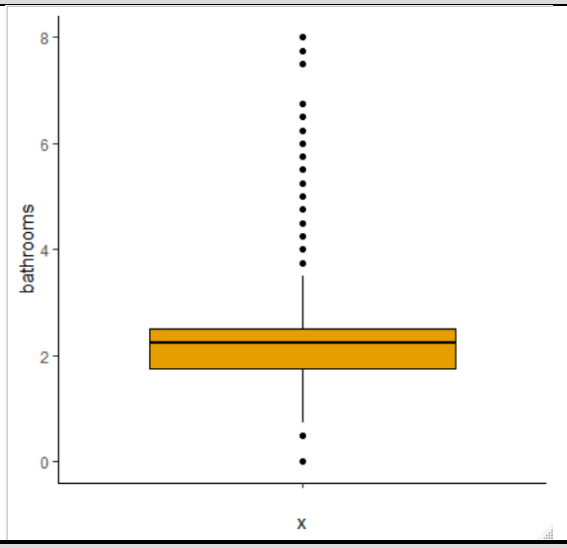
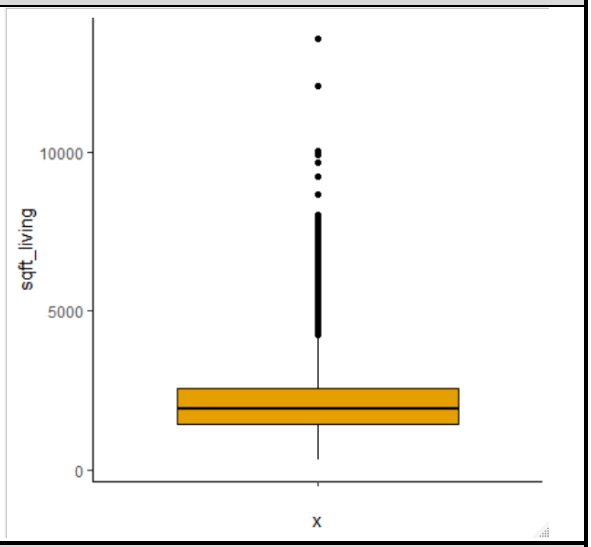
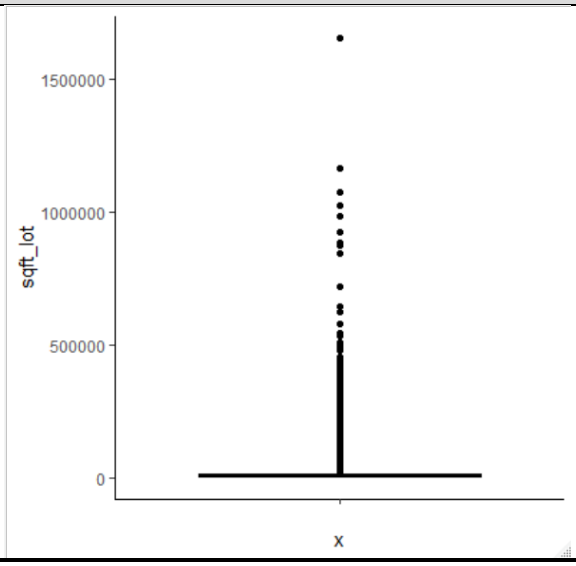
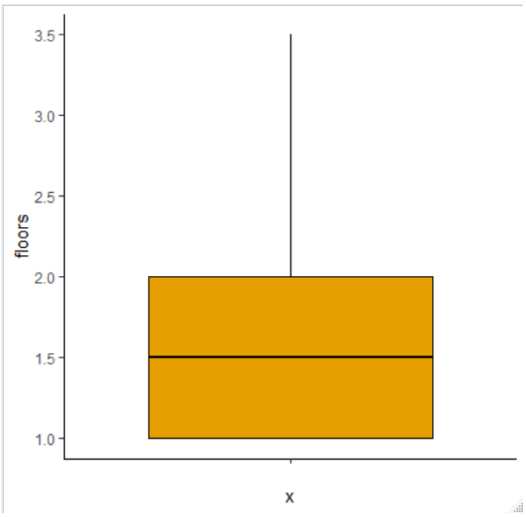
Q1. 모델 구축에 필요하지 않은 변수

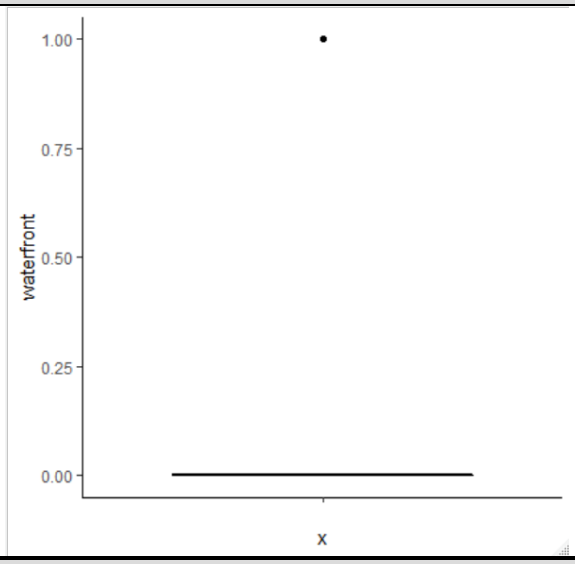
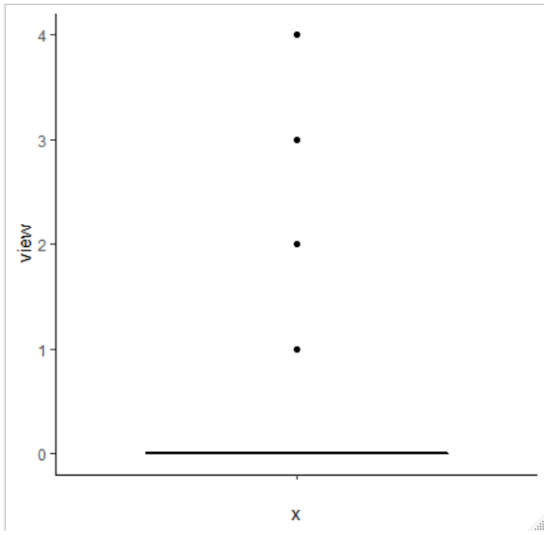
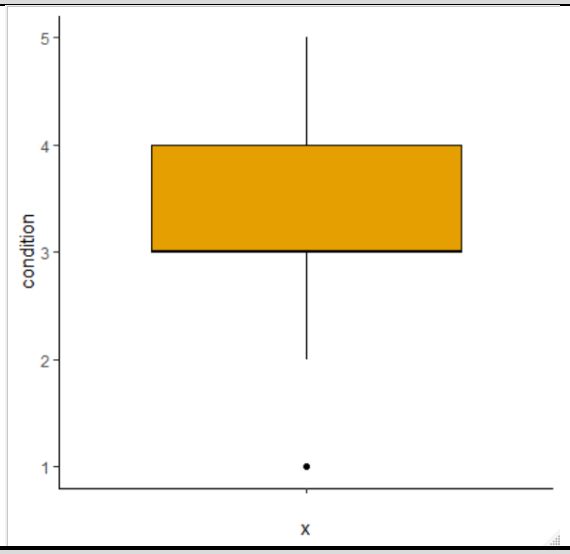
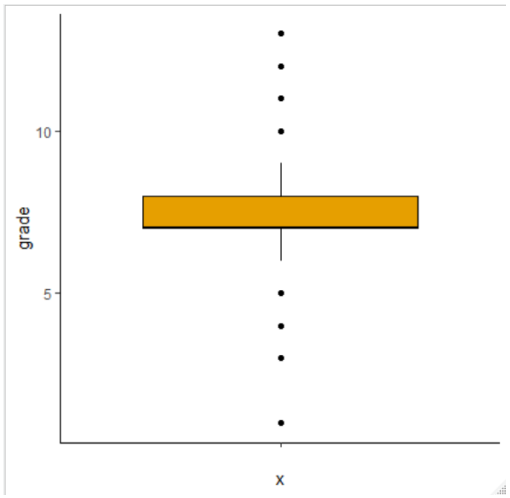
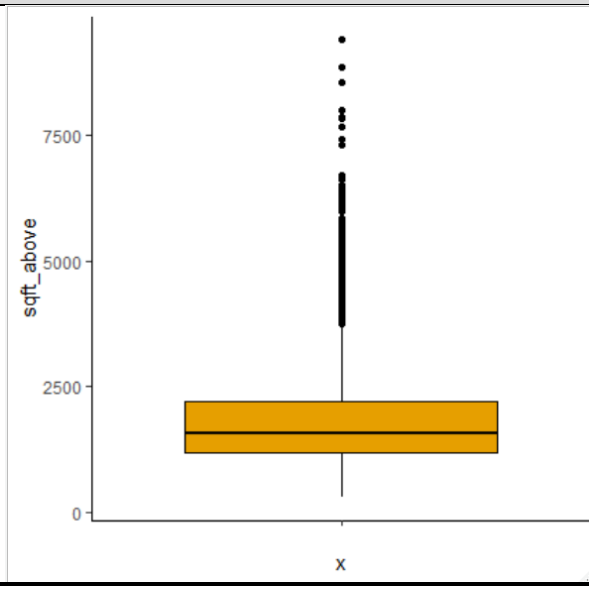
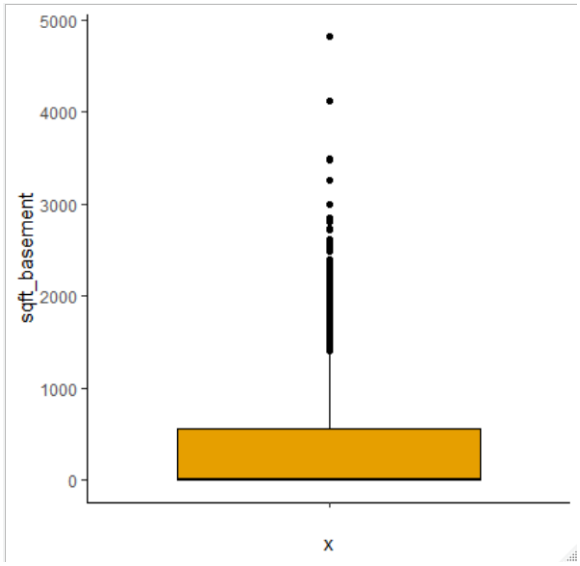
Id, date, zipcode는 MLR 모형 구축을 위해 필요하지 않습니다. ID는 서로 다른 집을 식별하는 코드일 뿐이고, date는 팔린 날짜 뒤에 T000000이 붙어있어 그대로 사용할 수 없습니다. 날짜 변수를 사용하려면 이를 년도, 월, 일을 구분하는 것이 좋다고 생각했습니다. Zipcode 역시 지역을 구분하는 코드로 모델 구축에 필요하지 않다고 생각했습니다.

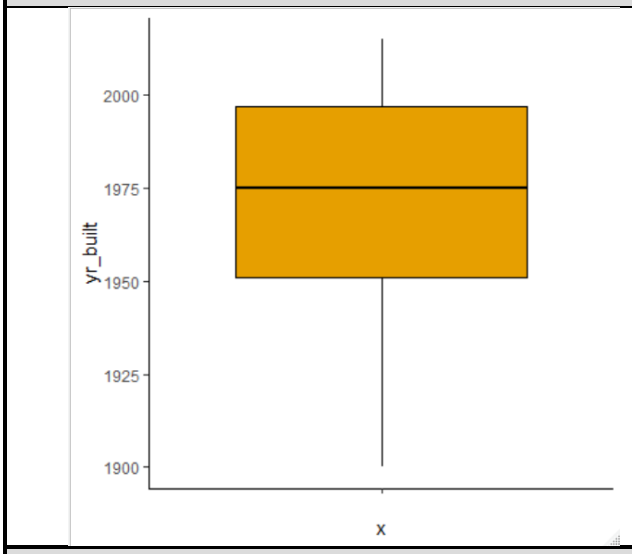
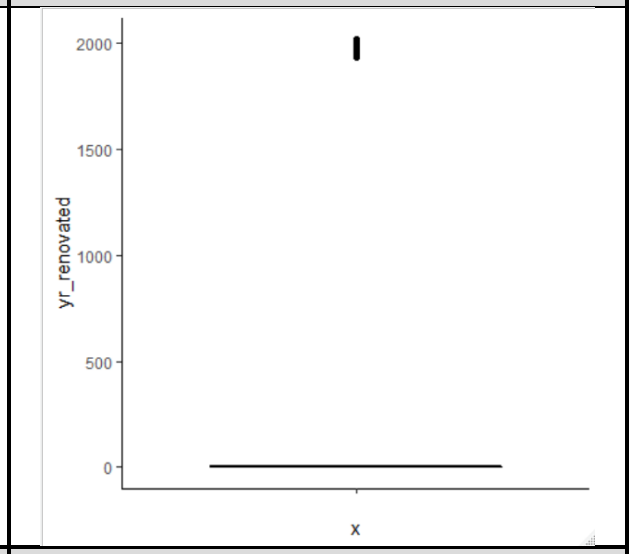
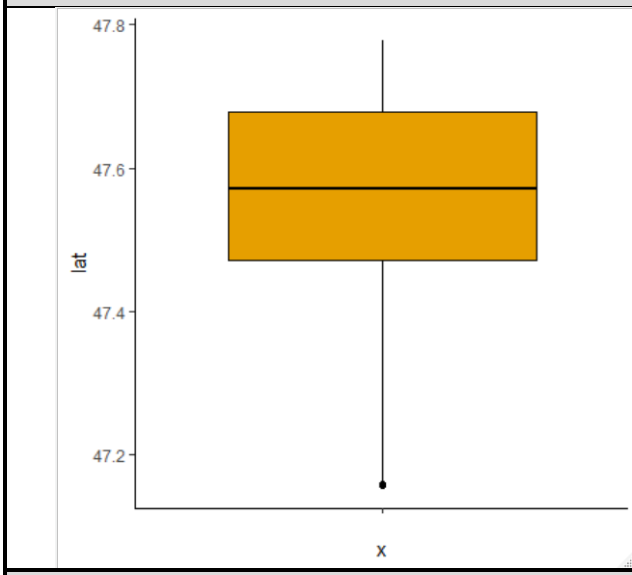
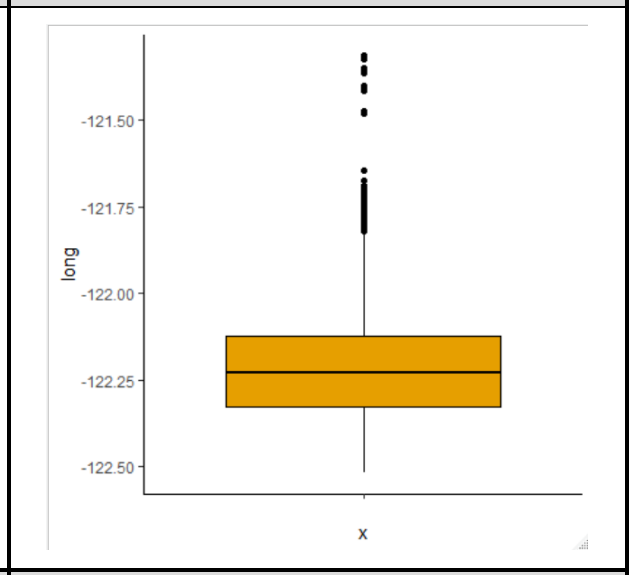
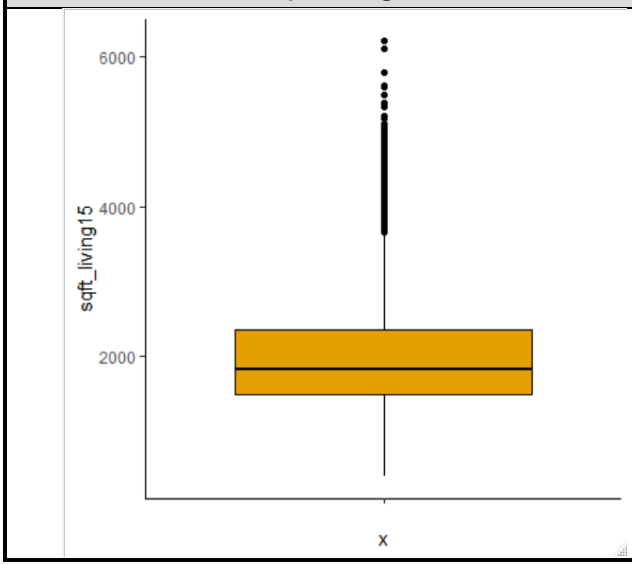
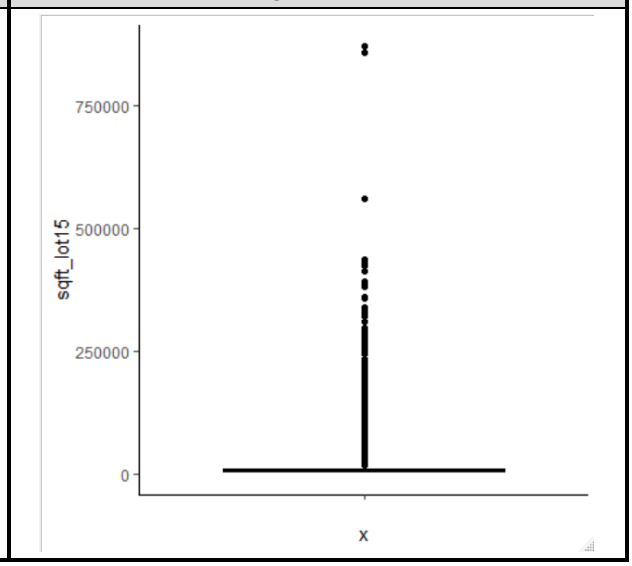
Q2. 입력 변수들의 단변량 통계량과 Boxplot

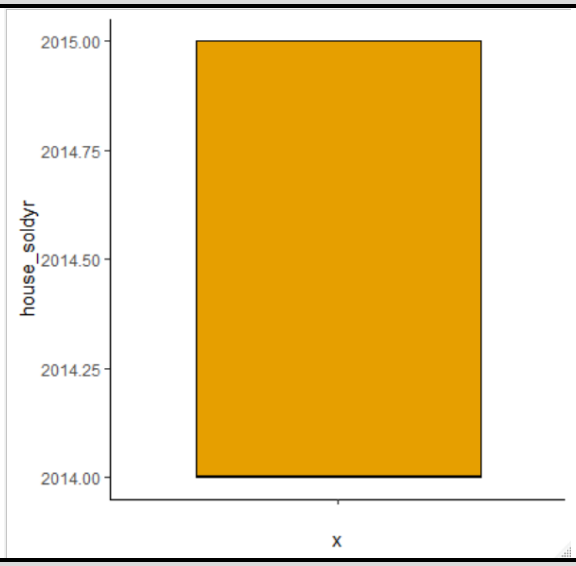
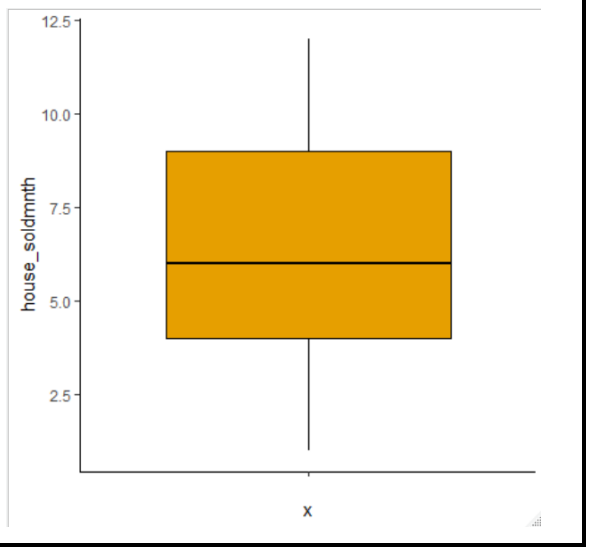
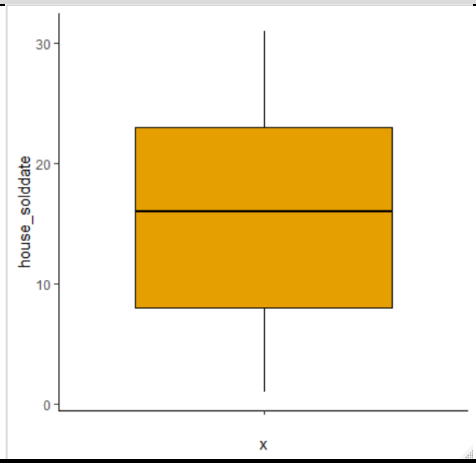
개별 입력 변수들에 대한 Mean, Standard Deviation, Skewness, Kurtosis와 Box plot은 다음과 같습니다. 기존의 date변수는 연, 월, 일로 쪼개 house_soldyr, house_soldmnth, house_solddate라는 새로운 변수에 각각 저장했습니다. 기타 위에서 필요하지 않다고 판단한 변수들은 제거했습니다.

	variable	mean	sd	skew	kurtosis
1	price	5.400881e+05	3.671272e+05	4.023510600	34.5737851
2	bedrooms	3.370842e+00	9.300618e-01	1.974025501	49.0472092
3	bathrooms	2.114757e+00	7.701632e-01	0.511036631	1.2789328
4	sqft_living	2.079900e+03	9.184409e+02	1.471351173	5.2408399
5	sqft_lot	1.510697e+04	4.142051e+04	13.058206214	284.9849447
6	floors	1.494309e+00	5.399889e-01	0.616091195	-0.4851211
7	waterfront	7.541757e-03	8.651720e-02	11.383527677	127.5906058
8	view	2.343034e-01	7.663176e-01	3.395278260	10.8889389
9	condition	3.409430e+00	6.507430e-01	1.032661283	0.5250381
10	grade	7.656873e+00	1.175459e+00	0.770996171	1.1899912
11	sqft_above	1.788391e+03	8.280910e+02	1.446463675	3.4006466
12	sqft_basement	2.915090e+02	4.425750e+02	1.577746032	2.7141396
13	yr_built	1.971005e+03	2.937341e+01	-0.469740189	-0.6577498
14	yr_renovated	8.440226e+01	4.016792e+02	4.548861892	18.6945407
15	lat	4.756005e+01	1.385637e-01	-0.485203120	-0.6766492
16	long	-1.222139e+02	1.408283e-01	0.884930137	1.0486058
17	sqft_living15	1.986552e+03	6.853913e+02	1.108027459	1.5960234
18	sqft_lot15	1.276846e+04	2.730418e+04	9.505423701	150.7137317
19	house_soldyr	2.014323e+03	4.676160e-01	0.757194017	-1.4267232
20	house_soldmnth	6.574423e+00	3.115308e+00	0.063121415	-1.0081480
21	house_solddate	1.568820e+01	8.635063e+00	-0.005676743	-1.1960669

Price	Bedrooms
	
Bathrooms	**Sqft_living**
	
Sqft_loft	**Floors**
	

Waterfront	View
	
Condition	**Grade**
	
Sqft_above	**Sqft_basement**
	

Yr_built	Yr_renovated
	
Lat	**long**
	
Sqft_living15	**Sqft_loft15**
	

House_soldyr	House_soldmonth
	
House_solddate	
	

전체 변수 중에서 정규분포를 따른다고 생각한 변수들은 다음과 같습니다.

(1) sqft-living: median을 나타내는 검정 선이 박스의 가운데 위치해서 정규분포의 양쪽이 대칭인 모양을 잘 나타낼 것 같았습니다. 위쪽으로 아웃라이어가 많이 표시되어 있으므로 오른쪽 꼬리가 조금 길 것 같습니다. Kurtosis도 5로 3에 가까운 편이고, skew도 1.47로 0에 가깝습니다.

(2) sqft_above: 역시 kurtosis가 3.40로 3에 가까웠고 skewness는 1.446으로 0에 가까웠습니다.

(3) long: Boxplot의 모양이 대칭이었지만 위쪽 아웃라이어가 많아 오른쪽 꼬리가 길 것으로 예상합니다. Kurtosis가 1.04, skewness는 0.88이었습니다.

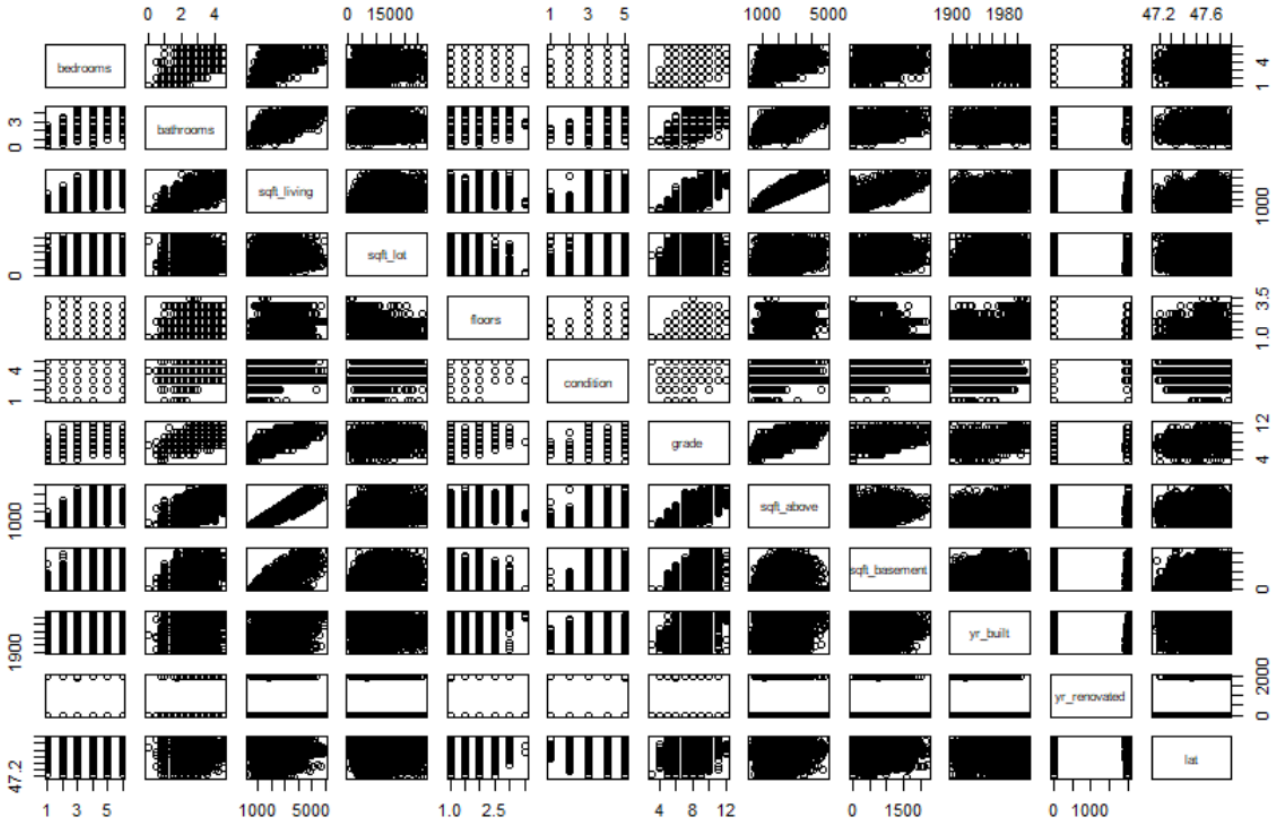
(4) sqft_living15: 역시 Boxplot의 모양이 비교적 대칭이었습니다. Kurtosis가 1.59, skewness는 1.10입니다.

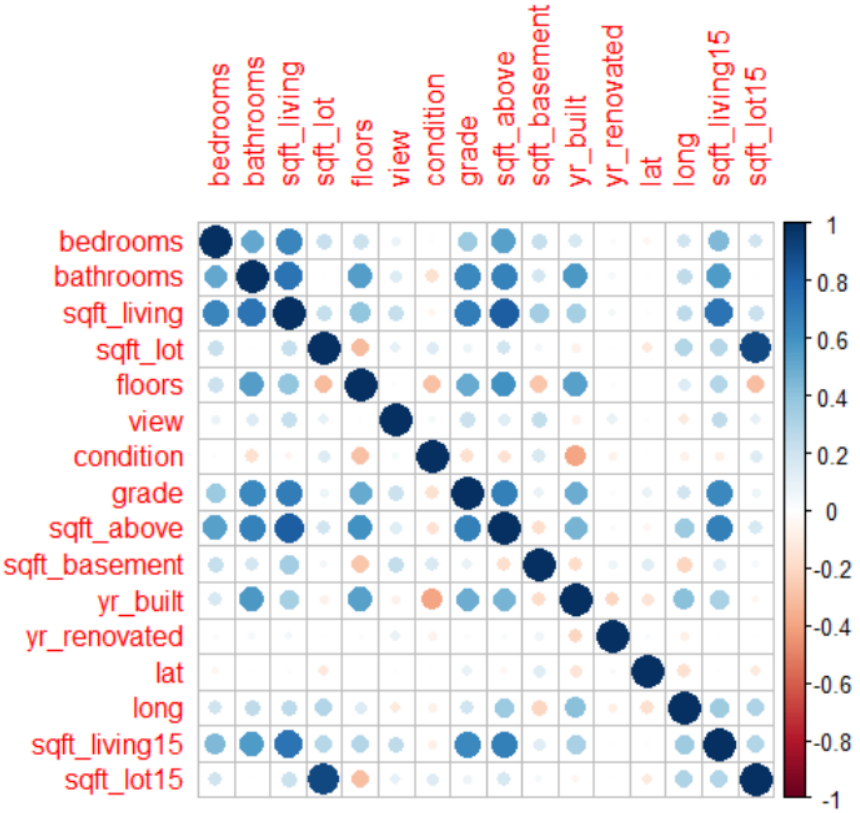
Q3. 이상치의 조건 정의와 제거

위의 Boxplot들을 봤을 때, 가운데 박스에서 점점 멀어질수록 이상치 값이 드물게 나타남을 알 수 있었습니다. 예를 들어, Sqft_above변수의 Boxplot에서 90% percentile보다 큰 값들이 꽤 가깝게 몰려 있고, 더 커질수록 이상치가 드물어졌습니다. 단순히 90% percentile를 초과하는 값을 이상치로 정의하기에는 제거되는 관측치 수가 너무 크다고 생각해 이상치를 [(Q1-3*IQR), (Q3+3*IQR)로 정의했습니다. 이에 해당하는 객체들을 데이터셋에서 제거했습니다. 단, waterfront와 같이 0과 1로 이루어진 binary변수, yr_built처럼 이상치가 존재하지 않는 변수는 이상치를 제거하지 않았습니다.

Q4. Scatterplot과 Correlation plot

수치형 변수들의 조합에 대해 산점도와 correlation plot을 그려보았습니다.





이로부터 sqft_lot15와 sqft_lot의 상관관계가 가장 높음을 알 수 있습니다. Sqft_lot15는 2015년의 lot size고, sqft_lot은 이전 lot size입니다. 리모델링을 하지 않았다면 lot size가 동일할 것이므로 이렇게 상관관계가 높게 나왔다고 생각합니다. 마찬가지로 Sqft_living15와 Sqft_living도 높은 상관관계가 있었습니다. Sqft_above와 Sqft_living역시 큰 상관관계가 있었습니다. Sqft_above는 지하를 제외한 집의 size이므로 livingroom size가 크면 sqft_above도 큰 값을 가지게 됩니다. Grade와 sqft_living도 큰 양의 상관관계가 있었습니다.

Q5. 학습 데이터로 MLR 모델 구축

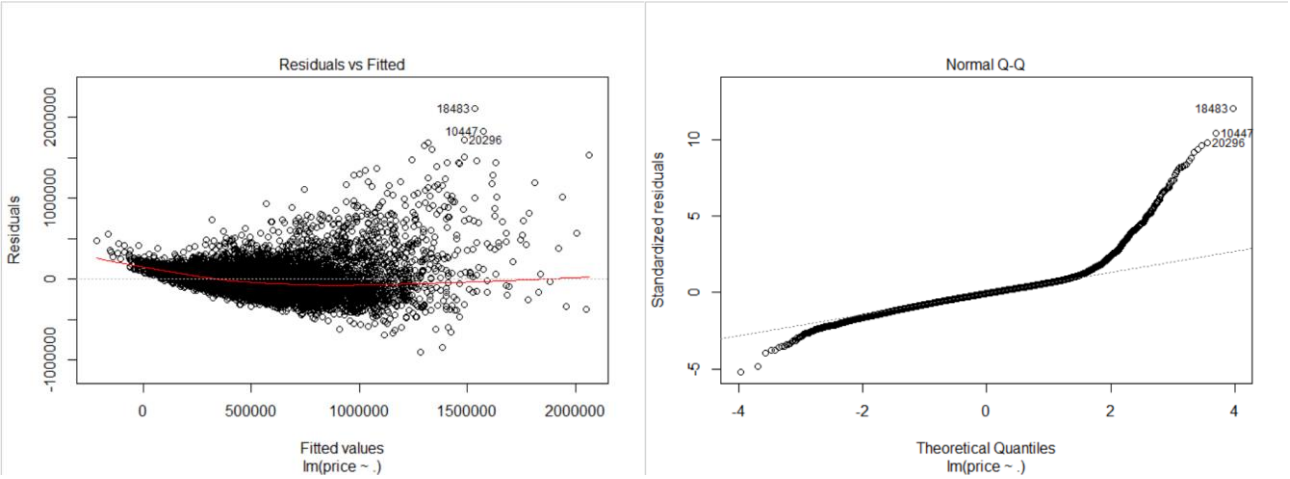
전체 데이터셋을 70%의 학습 데이터와 30%의 테스트 데이터로 분할한 후 모든 변수를 사용하여 MLR 모델을 학습했습니다. 결과는 다음과 같습니다.

```
> summary(mlr_house)

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

Residuals:
    Min      1Q  Median      3Q     Max
-904307  -93983   -9521   70934 2115481

Coefficients: (1 not defined because of singularities)
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -3.303e+07  1.885e+06 -17.522  < 2e-16 ***
bedrooms       -2.941e+04  2.316e+03 -12.696  < 2e-16 ***
bathrooms       3.409e+04  3.771e+03   9.039  < 2e-16 ***
sqft_living     1.183e+02  5.293e+00  22.359  < 2e-16 ***
```

```
sqft_lot         -2.433e+00  7.184e-01  -3.386 0.000711 ***
floors            9.910e+03  4.388e+03   2.258 0.023942 *
waterfront1       4.475e+05  2.316e+04  19.321  < 2e-16 ***
view              4.730e+04  2.484e+03  19.045  < 2e-16 ***
condition         3.380e+04  2.587e+03  13.063  < 2e-16 ***
grade             1.050e+05  2.471e+03  42.486  < 2e-16 ***
sqft_above        2.800e+01  5.239e+00   5.346 9.16e-08 ***
sqft_basement            NA         NA      NA       NA
yr_built         -2.475e+03  7.963e+01 -31.082  < 2e-16 ***
yr_renovated      3.254e+01  4.176e+00   7.792 7.05e-15 ***
lat               5.707e+05  1.189e+04  47.993  < 2e-16 ***
long             -8.153e+04  1.432e+04  -5.692 1.28e-08 ***
sqft_living15     4.934e+01  4.170e+00  11.832  < 2e-16 ***
sqft_lot15        1.267e+00  8.479e-01   1.494 0.135199
house_soldyr2015  5.605e+04  1.068e+04   5.248 1.56e-07 ***
house_soldmnth2   1.951e+04  9.780e+03   1.995 0.046058 *
house_soldmnth3   3.141e+04  9.056e+03   3.469 0.000525 ***
...(생략)
house_solddate31  1.416e+03  1.700e+04   0.083 0.933604
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 176200 on 13520 degrees of freedom
Multiple R-squared:  0.6962,    Adjusted R-squared:  0.6949
F-statistic: 534.2 on 58 and 13520 DF,  p-value: < 2.2e-16
```

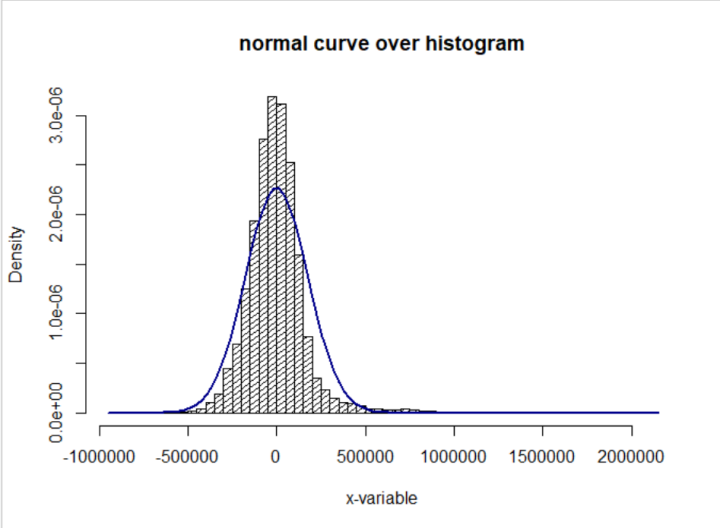
Adjusted R2값은 0.6949였습니다. 이는 데이터가 비교적 선형성을 띄고 있고, 모델 구축에 사용한 변수들로 전체 변동의 69%를 설명할 수 있음을 나타냅니다.

Residual plot과 Q-Q Plot은 다음과 같습니다. Ordinary Least Square 방식의 솔루션은 residual이 independent하며 평균이 0이고 분산이 constant한 정규분포를 따라야 한다는 가정이 있습니다.



Residual Plot을 보았을 때, 잔차가 특정한 패턴 없이 골고루 퍼져있어야 합니다. 하지만 이 plot의 경우, 잔차가 깔때기 모양으로 점점 퍼져 나가는 것을 볼 수 있습니다. 따라서 잔차가 독립이라는 가정을 만족하기 힘들 것으로 보입니다. Normal QQ plot으로부터는 정규성을 확인할 수 있습니다. 비록 오른쪽 끝 부분에서 점들이 y=x 선과 많이 멀어지기는 했으나, 2의 근처에서 멀어지기 시작했으므로 대부분 정규성을 만족한다고 할 수 있습니다.

실제 그래프를 그려본 결과는 다음과 같습니다. 정규분포처럼 bell-shape이지만 정규분포보다 뾰족함을 알 수 있습니다.



Q6. 유의미한 변수들

유의수준 0.01에서 모형 구축에 통계적으로 유의미한 변수들을 파악하기 위해 개별 변수의 p-value를 살펴보았습니다. P-value가 0.01보다 낮다는 것은 coefficient의 값이 0이라는 귀무가설을 기각할 수 있는 것이므로 변수가 유의미함을 나타내기 때문입니다. P-value가 0.01보다 낮은 변수들은 다음과 같았습니다. 변수들의 coefficient 값이 양수라면 해당 변수가 한 단위 증가할 때, price가 coefficient만큼 증가한다는 것이므로 양의 상관관계를 갖고, 음수라면 한 단위 증가할 때마다 price가 그만큼 감소하므로 음의 상관관계를 갖습니다. 양의 상관관계를 갖는 변수들은 bathrooms, sqft_living, waterfront 등이었고, 음의 상관관계를 갖는 변수들은 bedrooms, sqft_loft 등이었습니다.

변수	Coefficient	P-value
bedrooms	-2.941e+04	< 2e-16v
bathrooms	3.409e+04	< 2e-16
sqft_living	1.183e+02	< 2e-16
sqft_lot	-2.433e+00	0.000711
waterfront1	4.475e+05	< 2e-16
view	4.730e+04	< 2e-16
condition	3.380e+04	< 2e-16
grade	1.050e+05	< 2e-16
sqft_above	2.800e+01	9.16e-08
yr_built	-2.475e+03	< 2e-16
yr_renovated	3.254e+01	7.05e-15
lat	5.707e+05	< 2e-16
long	-8.153e+04	1.28e-08
sqft_living15	4.934e+01	< 2e-16
house_soldyr2015	5.605e+04	1.56e-07

Q7. Test 데이터셋의 MAE, MAPE, RMSE

Test 데이터셋에 대하여 MAE, MAPE, RMSE를 계산했습니다.

	RMSE	MAE	MAPE
kc_house	168912.4	113890.7	24.40328

MAE는 절대평균오차로 실제값과 예측값의 차이의 절댓값의 평균입니다. 이 모델의 MAE는 113890.7로 평균적으로 이만큼의 차이가 있었음을 알 수 있습니다. MAE는 차이의 크기는 제공하지만 y의 스케일에 상관없이 계산됩니다. 이를 보완한 것이 MAPE로 이 모델의 값은 24.40328이었습니다. 이는 y값에 비해 얼마나 차이가 있었는지를 나타내는 것입니다. 마지막으로 RMSE는 168912.4로, 차이의 제곱의 평균에 루트를 씌운 것입니다. 부호의 영향을 제거하기 위해 제곱을 했기 때문에 MAE보다 큰 값이 계산되었습니다.

Q8. 7개 변수 선택

만약 7개의 입력 변수만을 사용하여 모델을 구축한다면 p-value가 낮은 순으로 고려하겠습니다. 선형회귀에서 만족해야 할 가정 중 하나는 변수들 간의 상관관계가 없다는 것입니다. 따라서 Q4에서 상관관계가 높은 변수들은 하나를 제거하고 하나만 선택하겠습니다.

우선 p-value만을 고려했을 때, 가장 낮은 값을 가진 bedrooms, bathrooms, sqft_living, waterfront1, view, condition, grade, yr_built, lat, sqft_living15가 선택되었습니다. 이 중에서 sqft_living과 sqft_living15는 높은 상관관계를 갖고 있습니다. Sqft_living의 t값이 22.359로 sqft_living15의 t값보다 2배 가량 크므로 sqft_living을 선택했습니다. Bathroom과 sqft_living도 마찬가지로 상관관계가 있는데, sqft_living의 t값이 더 크므로 bathroom을 제거하고 sqft_living을 선택했습니다. Bedroom도 마찬가지 이유로 제거되었습니다. 결과적으로 선택된 7개의 변수는 sqft_living, waterfront1, view, condition, grade, yr_built, lat입니다.

Q9. 7개 변수로 구축한 MLR 모형의 Performance measures

위에서 선택한 변수들만을 사용해 MLR 모형을 다시 학습했습니다. 결과는 아래와 같습니다.

```
Call:
lm(formula = price ~ ., data = house_trn_data2)

Residuals:
    Min      1Q  Median      3Q     Max
-957882  -94052  -11660   72719 2174190

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.407e+07  5.937e+05  -40.54   <2e-16 ***
sqft_living  1.463e+02  2.824e+00   51.81   <2e-16 ***
waterfront1  5.643e+05  2.308e+04   24.45   <2e-16 ***
view         5.157e+04  2.384e+03   21.63   <2e-16 ***
condition    2.678e+04  2.513e+03   10.66   <2e-16 ***
grade        1.221e+05  2.262e+03   53.97   <2e-16 ***
yr_built    -2.236e+03  6.302e+01  -35.48   <2e-16 ***
lat          5.818e+05  1.163e+04   50.04   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 176900 on 13571 degrees of freedom
Multiple R-squared:  0.6855,    Adjusted R-squared:  0.6854
F-statistic:  4227 on 7 and 13571 DF,  p-value: < 2.2e-16
```

Adjusted R2값은 0.6854였습니다. MAE, MAPE, RMSE는 다음과 같습니다.

	RMSE	MAE	MAPE
kc_house	180146.1	117464.5	24.90673

위에서 모든 변수를 사용해 모델을 구축했을 때와 비교해보면, Adjusted R2는 큰 차이가 없었습니다. 모든 변수 사용시에는 0.6949였고, 7개의 변수만을 사용했을 때에는 0.6854였습니다. 이는 유의미한 변수들만을 선택한다면 y를 충분히 설명할 수 있다는 것을 나타냅니다.

RMSE와 MAE, MAPE값도 마찬가지로 큰 차이가 없었습니다. 이 역시 위에서 선택한 7개의 변수들이 모든 변수들을 사용했을 때와 비슷하게 y를 잘 예측함을 나타냅니다.