



집값 예측모형구축

김영석 김은태

Contents



01 분석 목적 및 데이터 소개

02 데이터 전처리

03 데이터 분석

04 해석



01 분석 목적 및 데이터 소개

02 데이터 전처리

03 데이터 분석

04 해석

분석 목적



미국 아이오와 주의
에임스(Ames) 도시의
집 값을 예측하는
모델을 구축해본다



	ID	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	...	SalePrice
1	1	60	RL	65	8450	Pave	NA	...	208500
2	2	20	RL	80	9600	Pave	NA	...	181500
3	3	60	RL	68	11250	Pave	NA	...	223500
4	4	70	RL	60	9550	Pave	NA	...	140000
...
1459	1459	20	RL	68	9717	Pave	NA	...	142125
1460	1460	20	RL	75	9937	Pave	NA	...	147500

변수 : 81개

관측치 : 1460개



지붕

RoofStyle
RoofMatl

외벽

Exterior1st
Exterior2nd
ExterQual
ExterCond

추가외벽

MasVnrArea
MasVnrType



지하

BsmtQual
BsmtCond
BsmtExposure
BsmtFinType1
BsmtFinSF1
BsmtFinType2
BsmtFinSF2
BsmtUnfSF
TotalBsmtSF

벽난로

Fireplaces
FireplaceQu

수영장

PoolArea
PoolQC

욕실

BsmtFullBath
BsmtHalfBath
FullBath
HalfBath

차고지

GarageType
GarageYrBlt
GarageFinish
GarageCars
GarageArea
GarageQual
GarageCond

주방

KitchenAbvGr
KitchenQual

면적

1stFlrSF
2ndFlrSF
LowQualFinSF
GrLivArea

Deck & Porch

WoodDeckSF
OpenPorchSF
EnclosedPorch
3SsnPorch
ScreenPorch

방의 개수

BedroomAbvGr
TotRmsAbvGrd

난방시설

Heating
HeatingQC
CentralAir

기타

PavedDrive
Fence
MiscFeature
MiscVal
Functional
Electrical
Foundation

부지(땅)에 관한 변수 / 기타 변수



땅

LotFrontage
LotArea
LandSlope
LotConfig
LotShape
LandContour

주변 지역과 접근성

MSSubClass
MSZoning
Utilities
Street
Alley
Neighborhood
Condition1,2

시간

MoSold
YrSold
YearBuilt
YearRemodAdd

판매정보

SaleType
SaleCondition

기타

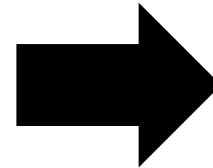
BldgType
HouseStyle
OverallQual
OverallCond



SalePrice (Target 변수)

범주형 변수 46개 (Street, Alley 등)

연속형 변수 34개 (LotArea 등)



회귀분석

Contents



01 데이터 소개 및 분석방법

02 데이터 전처리

03 데이터 분석

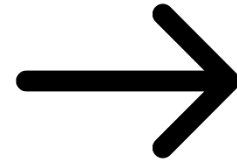
04 해석



결측치 처리

✓ 범주형변수

NA

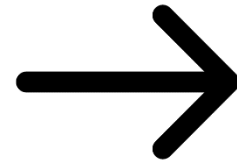


None

최빈값

✓ 수치형변수

NA



0



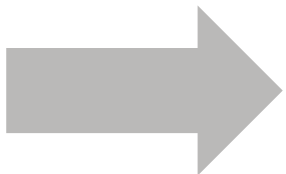
변수 제거

1차

범주형 변수 중 한 쪽으로 몰려있는 변수 확인

2차

SalePrice를 q4개의 범주로 나누어 독립성 검정을 실시



*Street Utilities LandSlope MiscFeature MiscVal Alley LandContour RoofMat1
BsmtCond Heating condition2* 총 11개의 변수제거



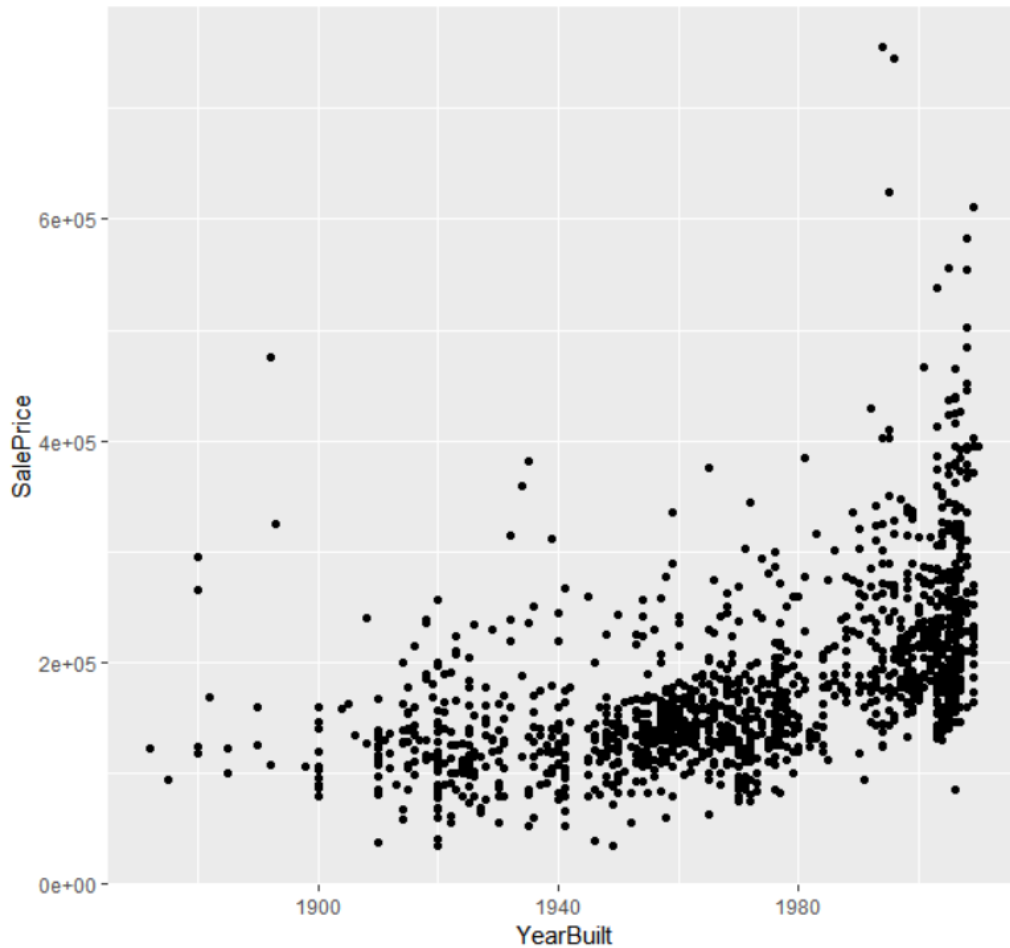
변수 제거

- 📎 YearBuilt(집이 지어진 년도) 는 YearRemodAdd(리모델링 년도) 변수로 설명가능
- 📎 Exterior2nd(외벽 재료2) 와 Exterior1st(외벽 재료1) 변수와 유사
- 📎 BldgType(주거 형태)는 KitchenAbvG(주방 개수) 변수로 설명가능
- 📎 YrSold, MoSold(매매 년도와 월) 특별한 패턴이 보이지 않음
- 📎 BedroomAbvGr(침실 개수) 는 TotRmsAbvGrd(방의 총 개수)로 대체

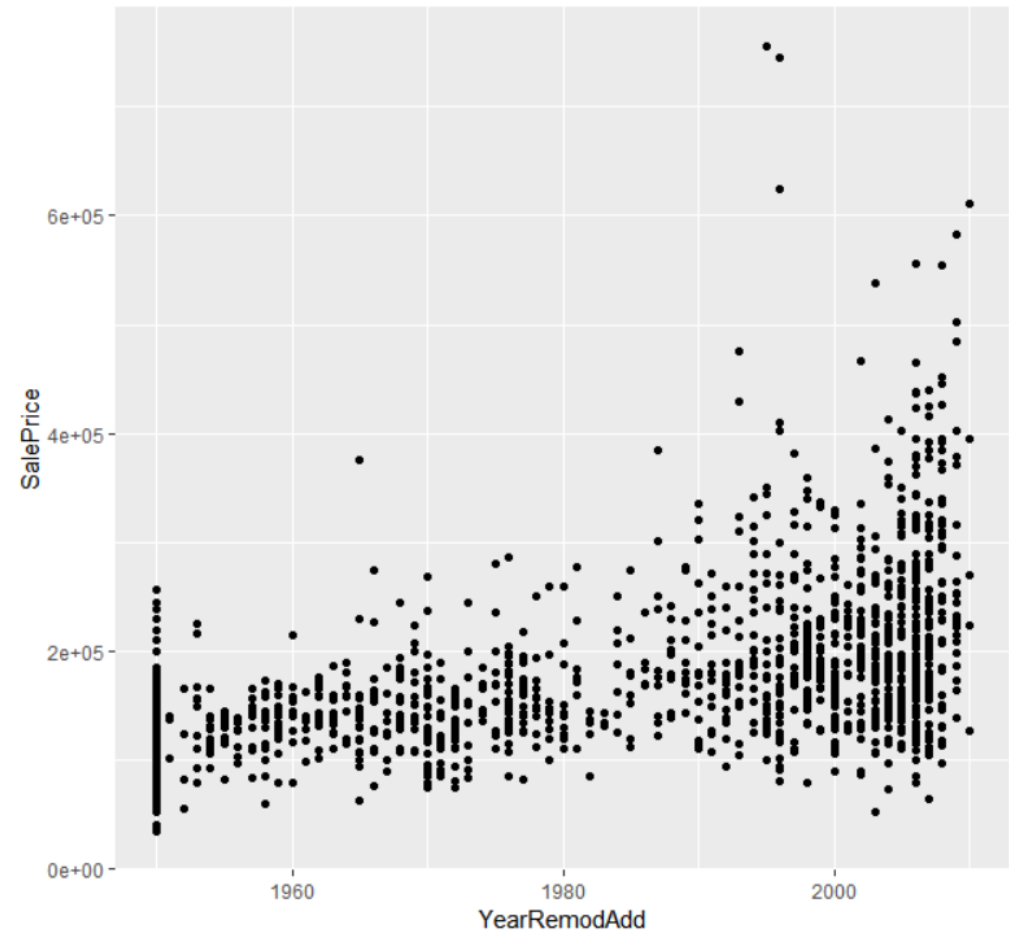


✓ *YearBuilt* 과 *YearRemodAdd*의 관계

건축년도 별 판매가격



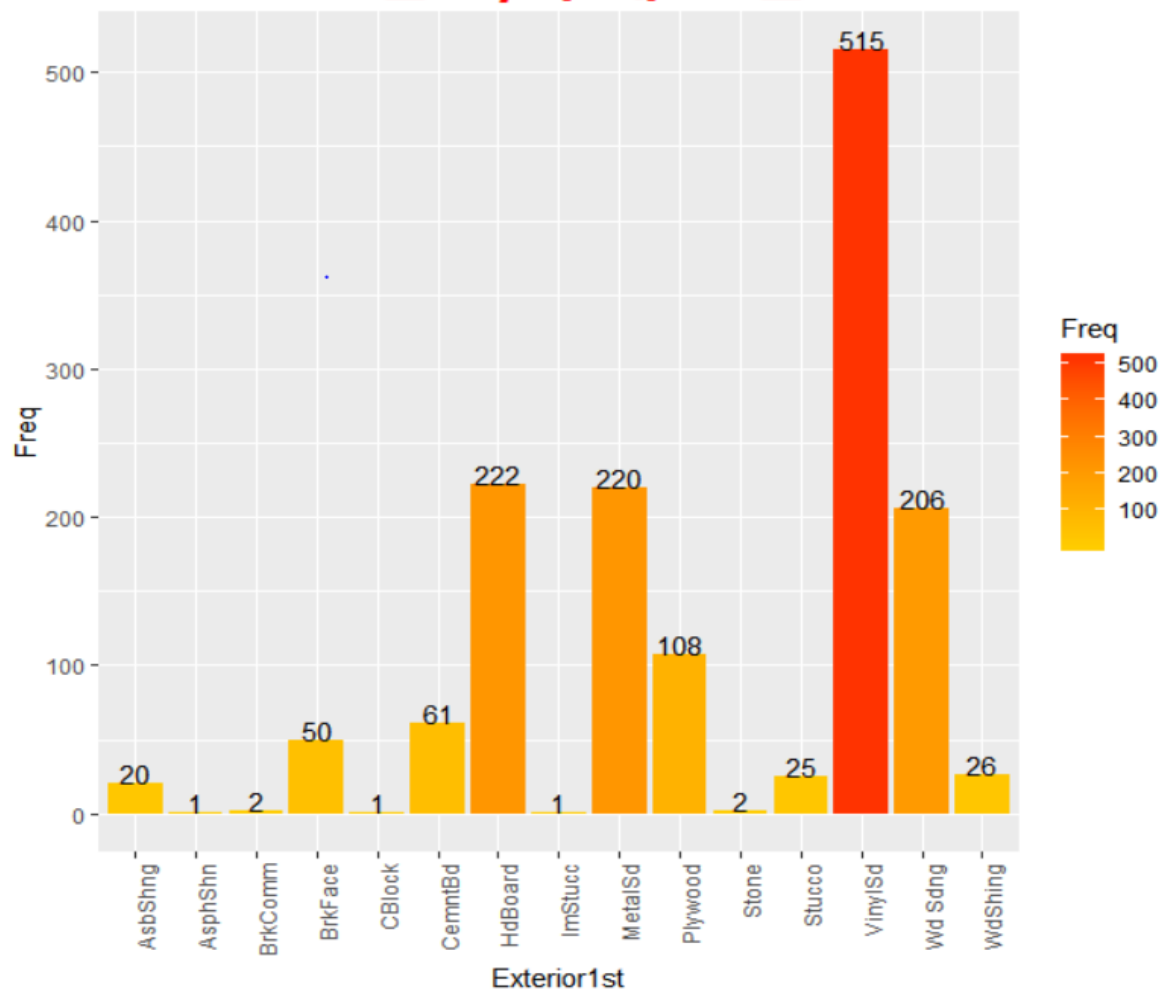
리모델링연도 별 판매가격



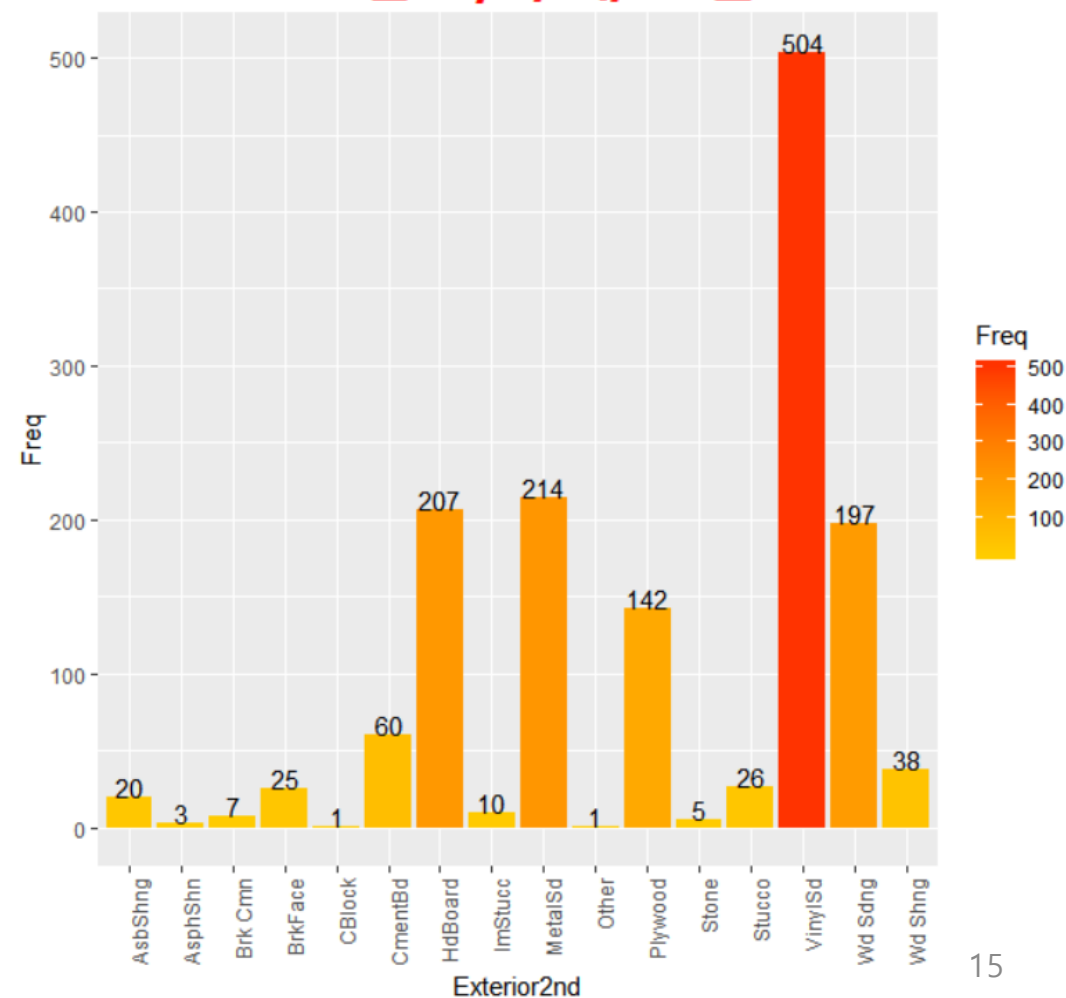


✓ Exterior 1st , Exterior 2nd

1st 집 외벽 재료 빈도



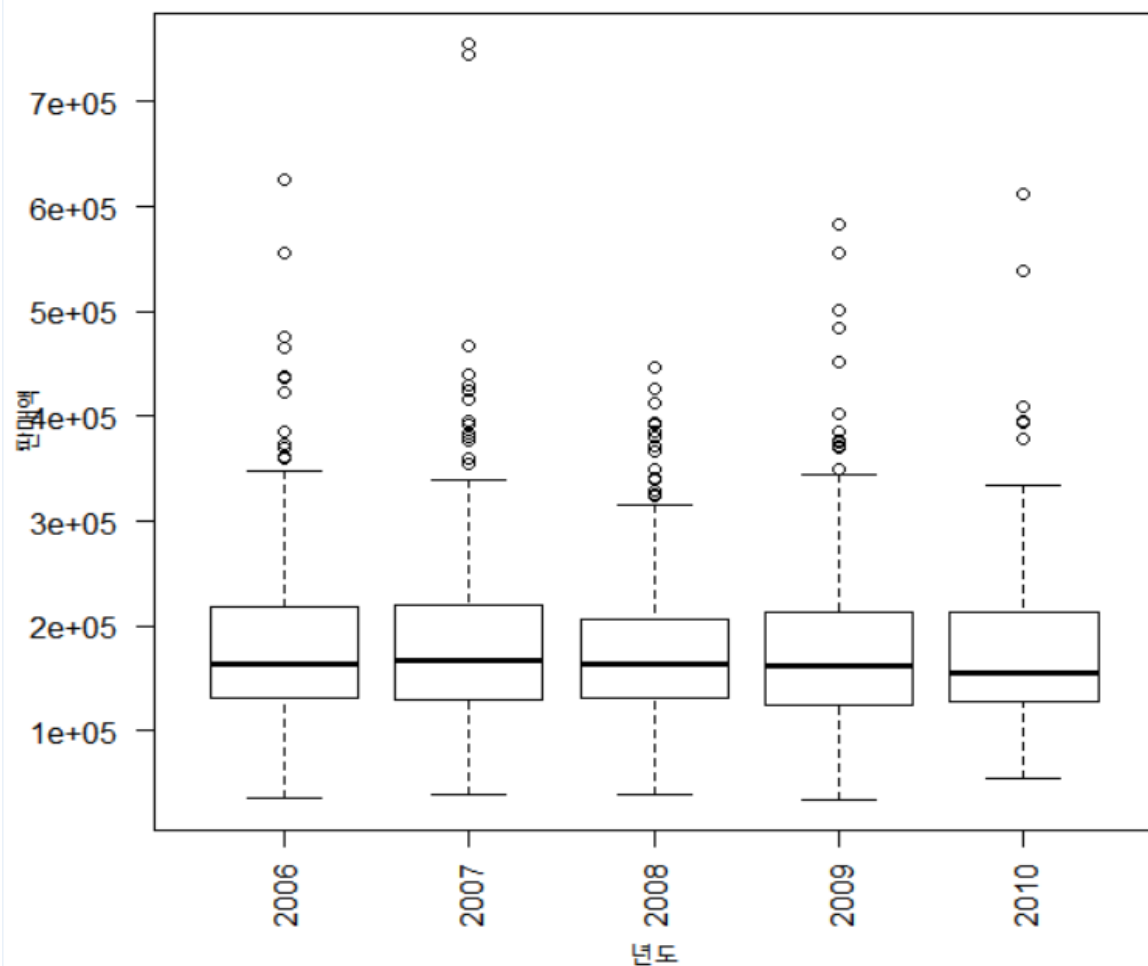
2nd 집 외벽 재료 빈도



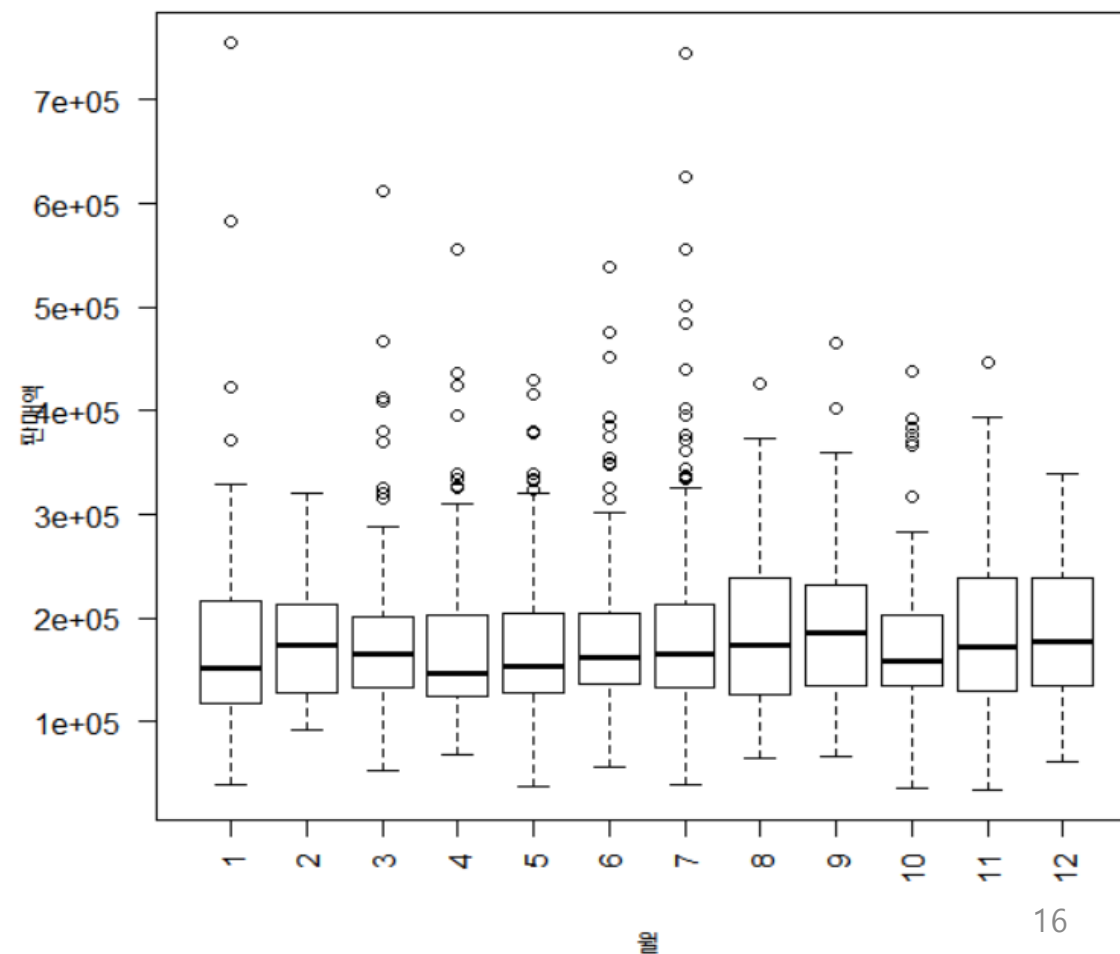


✓ YrSold(매매 년도) Mosold(매매 월) 과 SalePrice와의 관계

매매 년도



매매 월



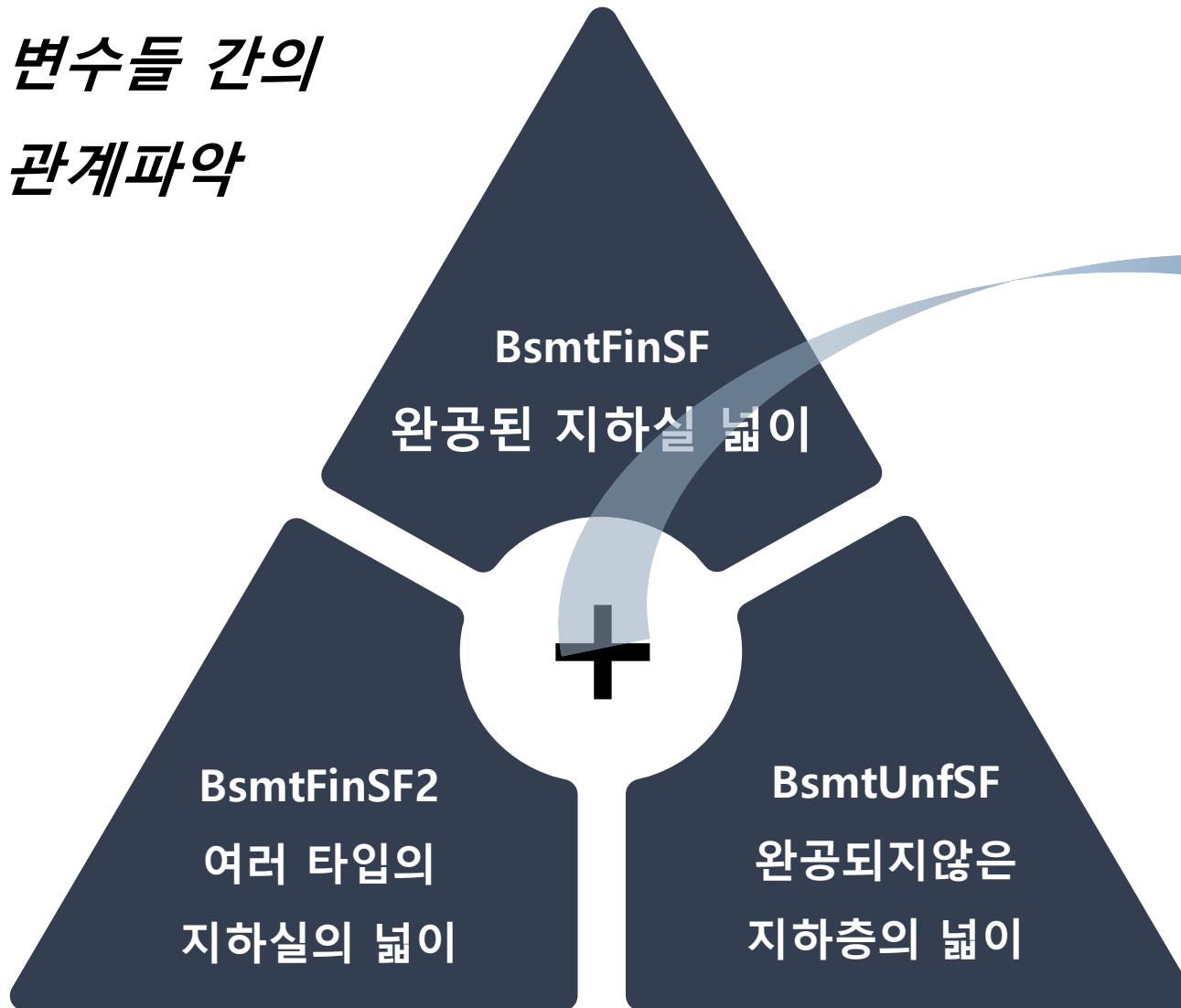


BldgType(주거형태)와 KitchenAbvGr(주방 개수)의 관계

주거형태 \ 주방 개수	1	2	3
1Fam	1214	5	1
2fmCon	14	16	1
Duplex	8	44	0
Twnhs	43	0	0
TwnhsE	114	0	0



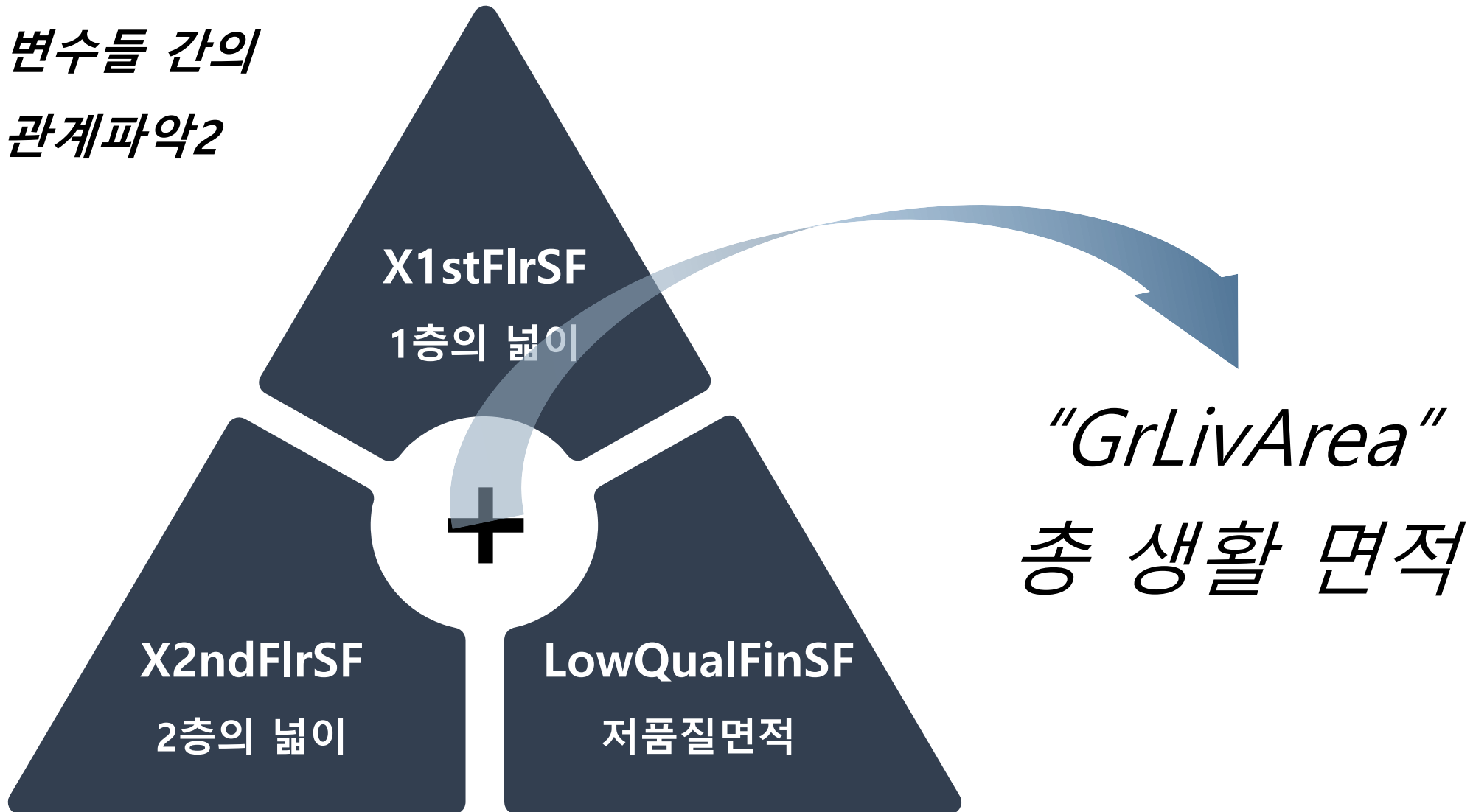
변수들 간의
관계파악



"TotalBsmtSF"
지하의 총 넓이

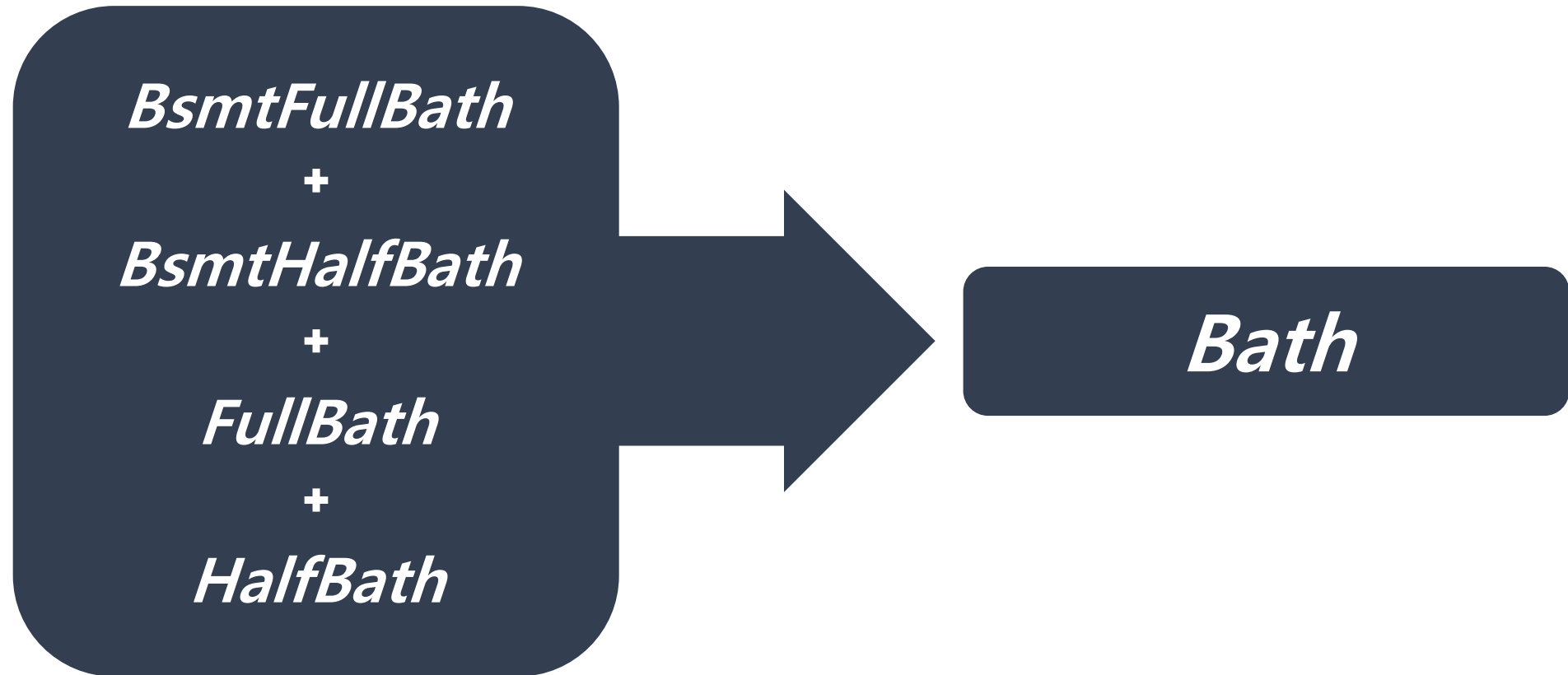


변수들 간의
관계파악2



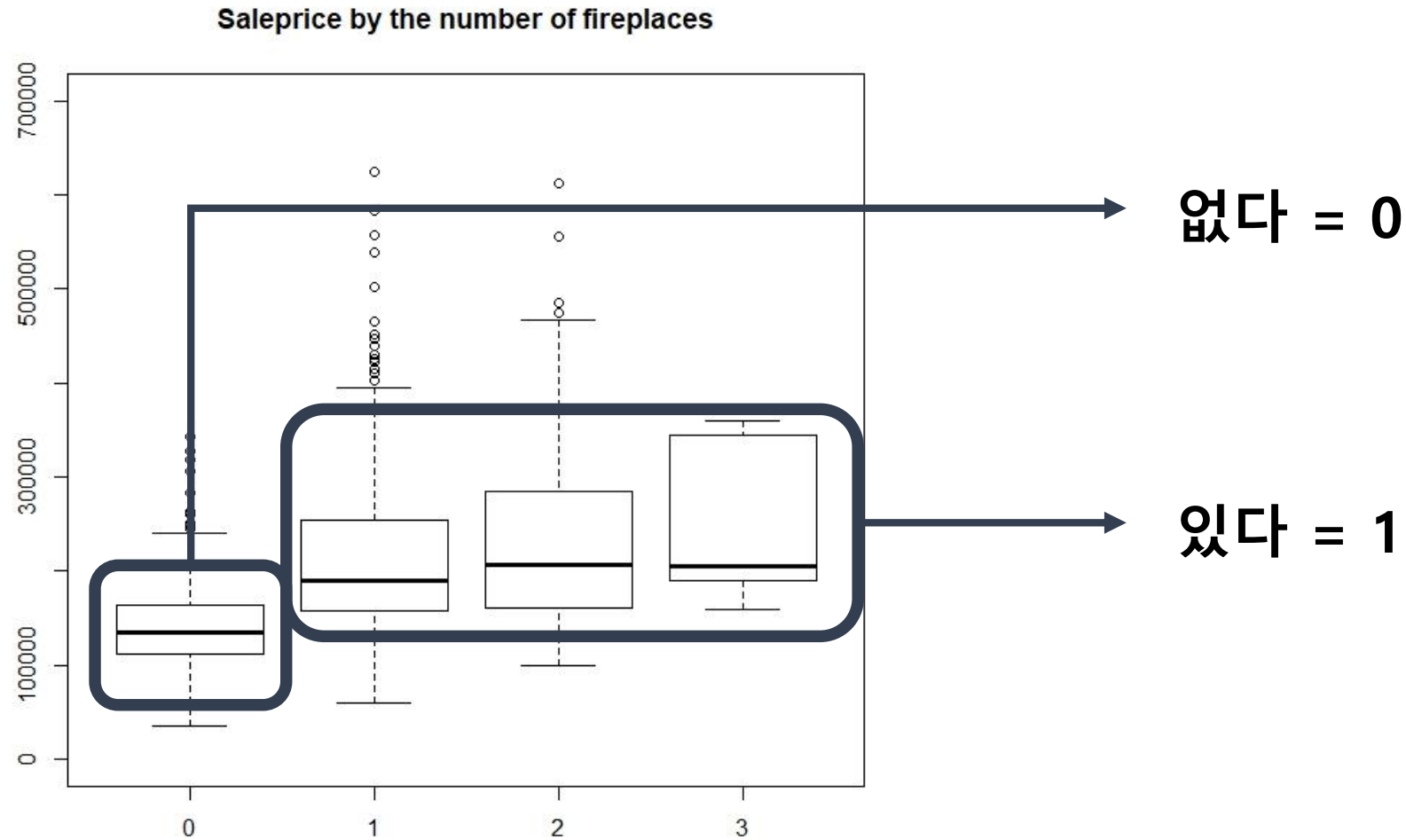


화장실 개수 **0000Bath**





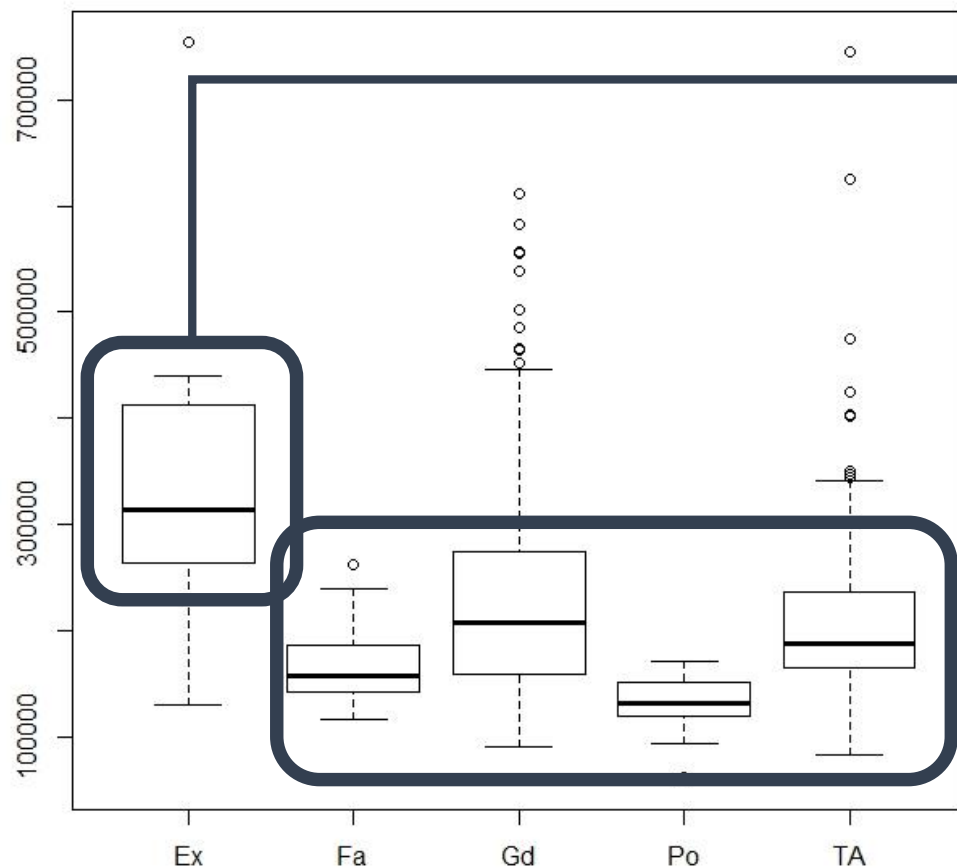
벽난로 개수 *Fireplaces*





벽난로 품질 FireplaceQu

Saleprice by the quality of fireplaces



좋다 = GOOD

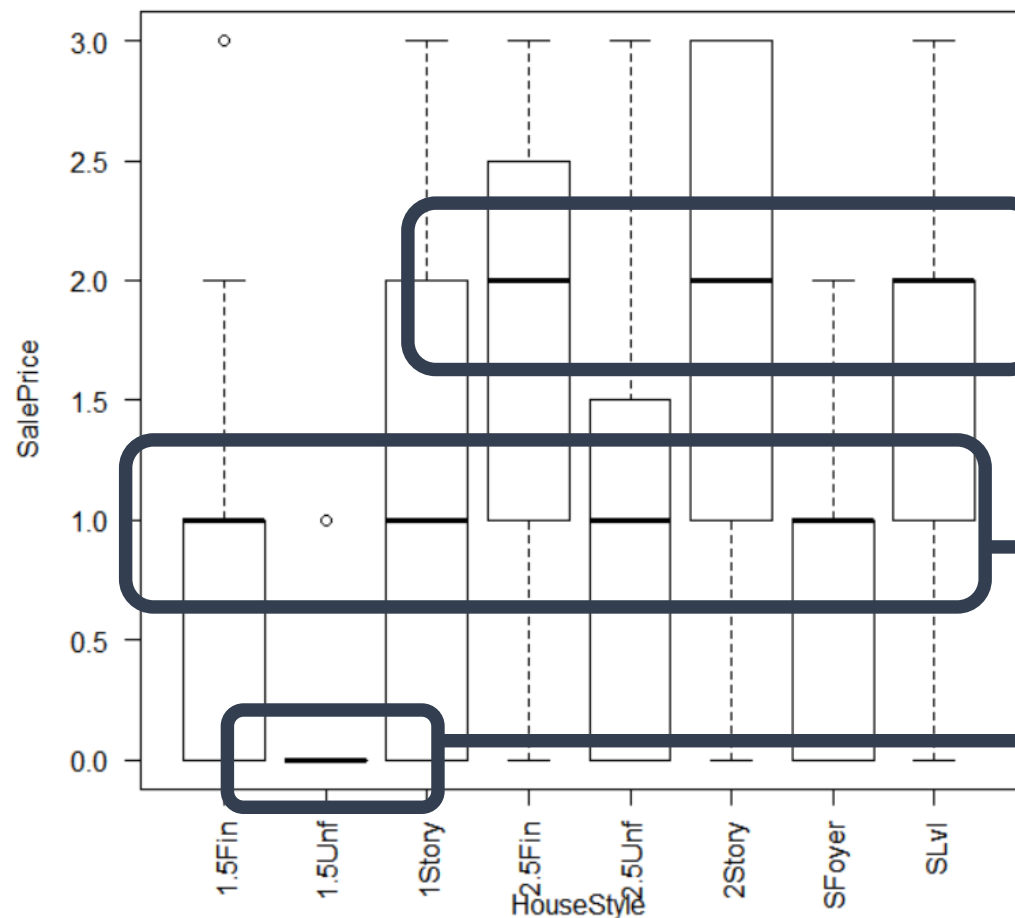
+ NA = "None"

나쁘다 = BAD



집의 형태 *HouseStyle*

HouseStyle boxplot



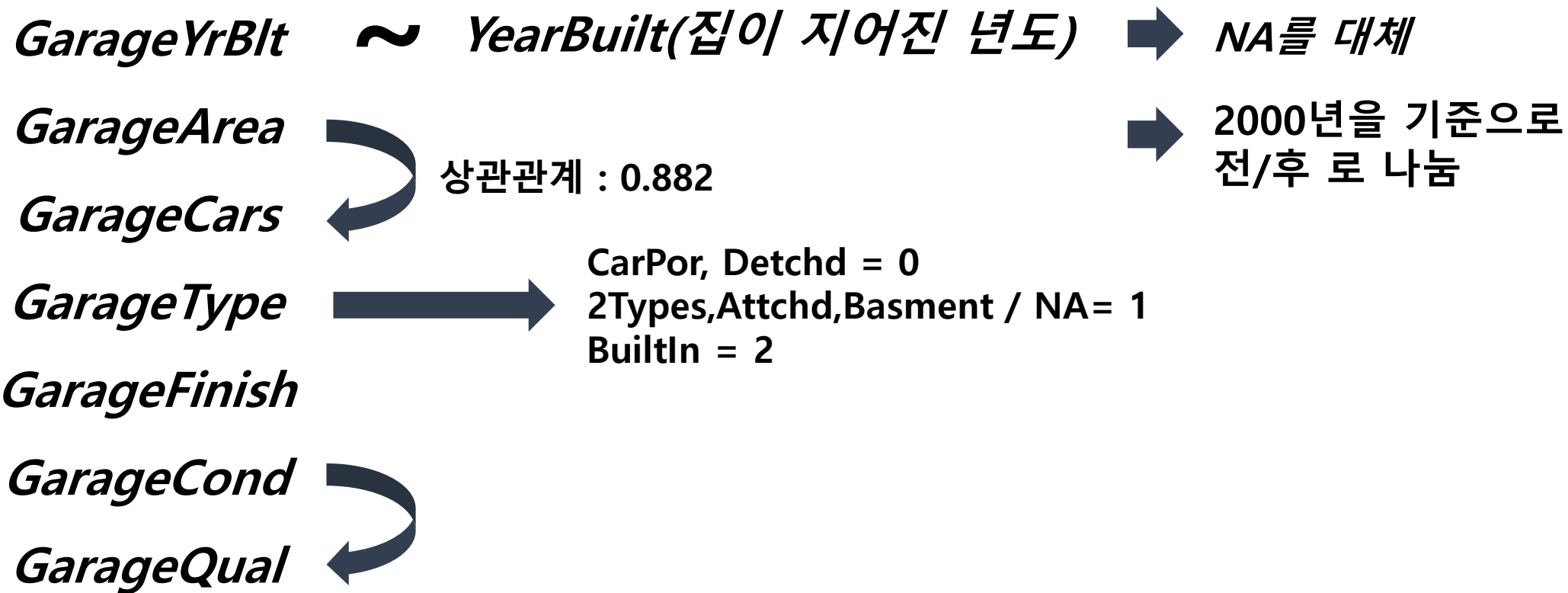
SalePrice 기준 평균 이용
2 부여

1 부여

0 부여



차고지 변수 **Garage**





그 외 변수 *Pool Fence WoodDeck ~Porch*



Pool : 수영장 면적



없다 = 0 / 있다 = 1



Fence : 울타리 면적



없다 = 0 / 있다 = 1



Wooddeck : Wooddeck 의 면적
~Porch : ~porch 의 면적



없다 = 0 / 있다 = 1



합친 후 *DeckPorch* 변수 생성

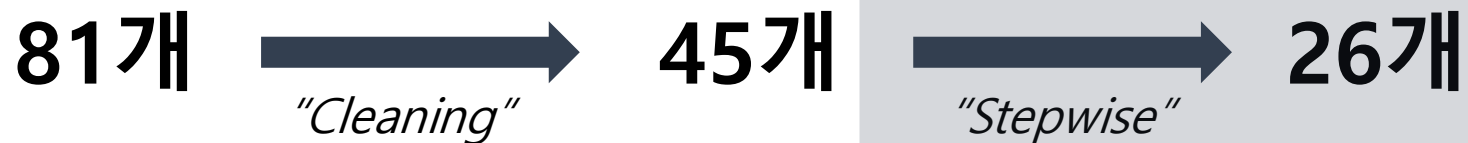


그 외 변수 *SaleType* *SaleCondition* *Functional*

✓ *SaleType* : 판매형태 ➡ TYPE1 / TYPE2

✓ *SaleCondition* : 판매조건 ➡ Normal / Partial

✓ *Functional* : 집의 기능 ➡ Maj2 / Typ



전처리 최종 변수

*LotArea LotShape Neighborhood Condition1 HouseStyle
OverallQual OverallCond RoofStyle MasVnrArea BsmtQual
BsmtExposure BsmtFinType1 TotalBsmtSF CentralAir GrLivArea
KitchenAbvGr KitchenQual TotRmsAbvGrd Fireplaces FireplaceQu
SaleType SaleCondition Bath GarageCars DeckPorch SalePrice*



01 분석 목적 및 데이터 소개

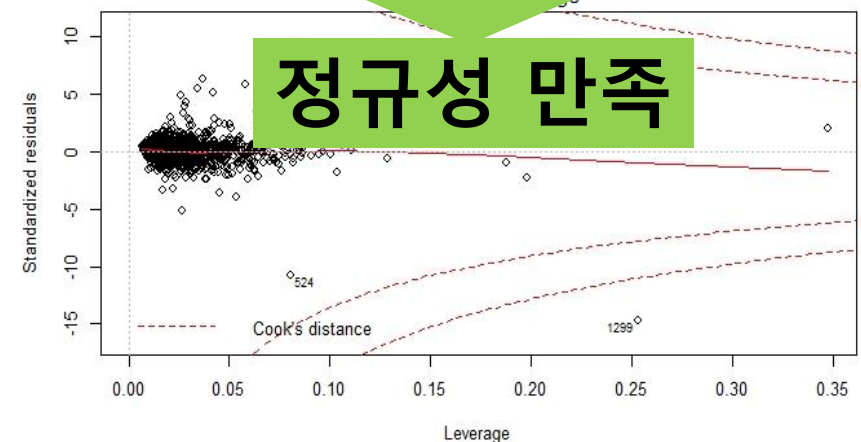
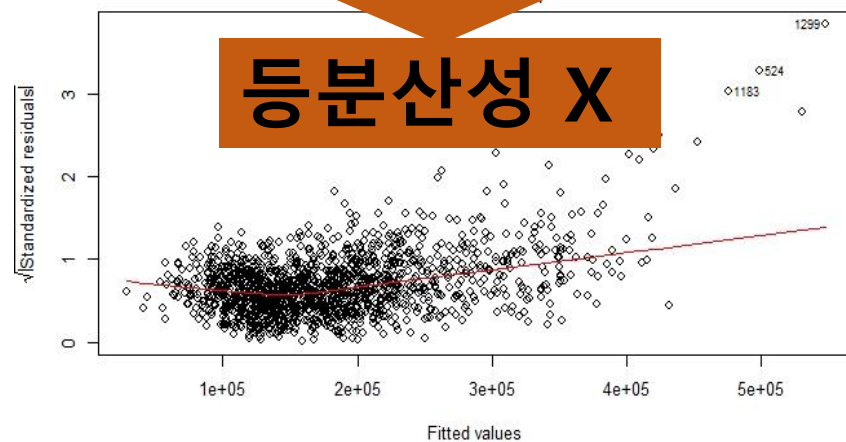
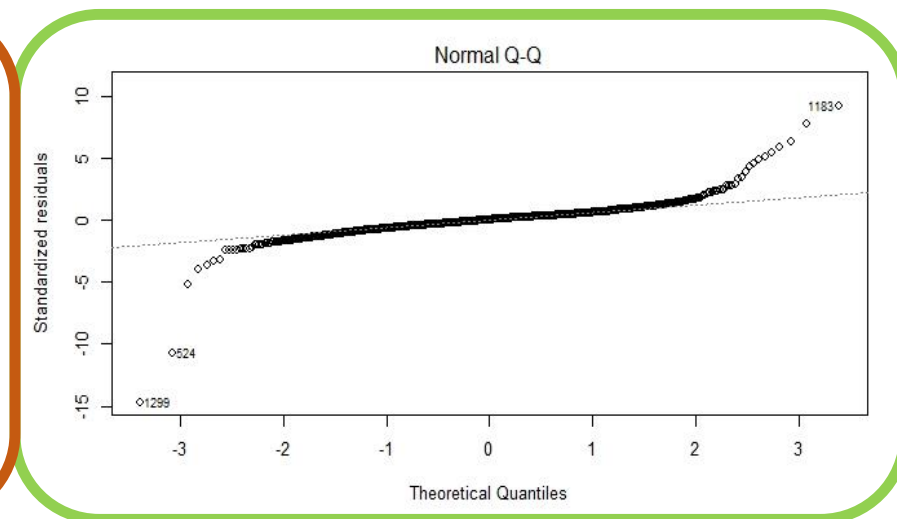
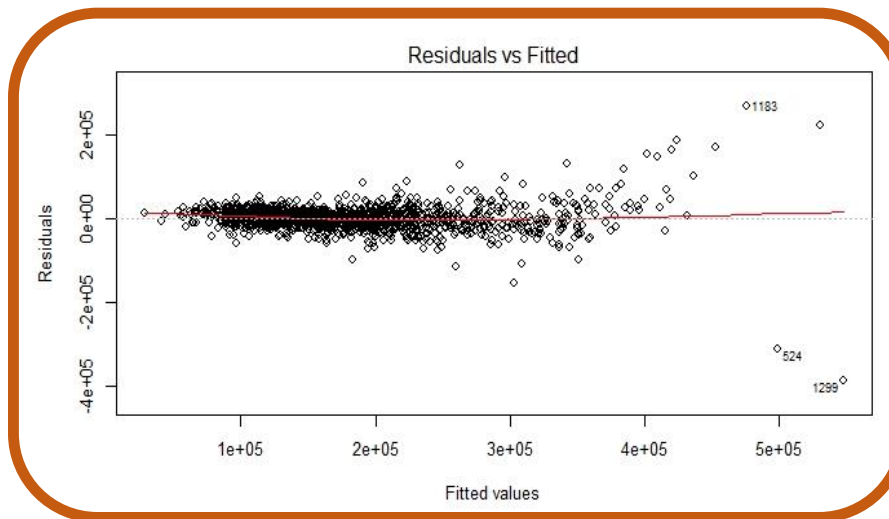
02 데이터 전처리

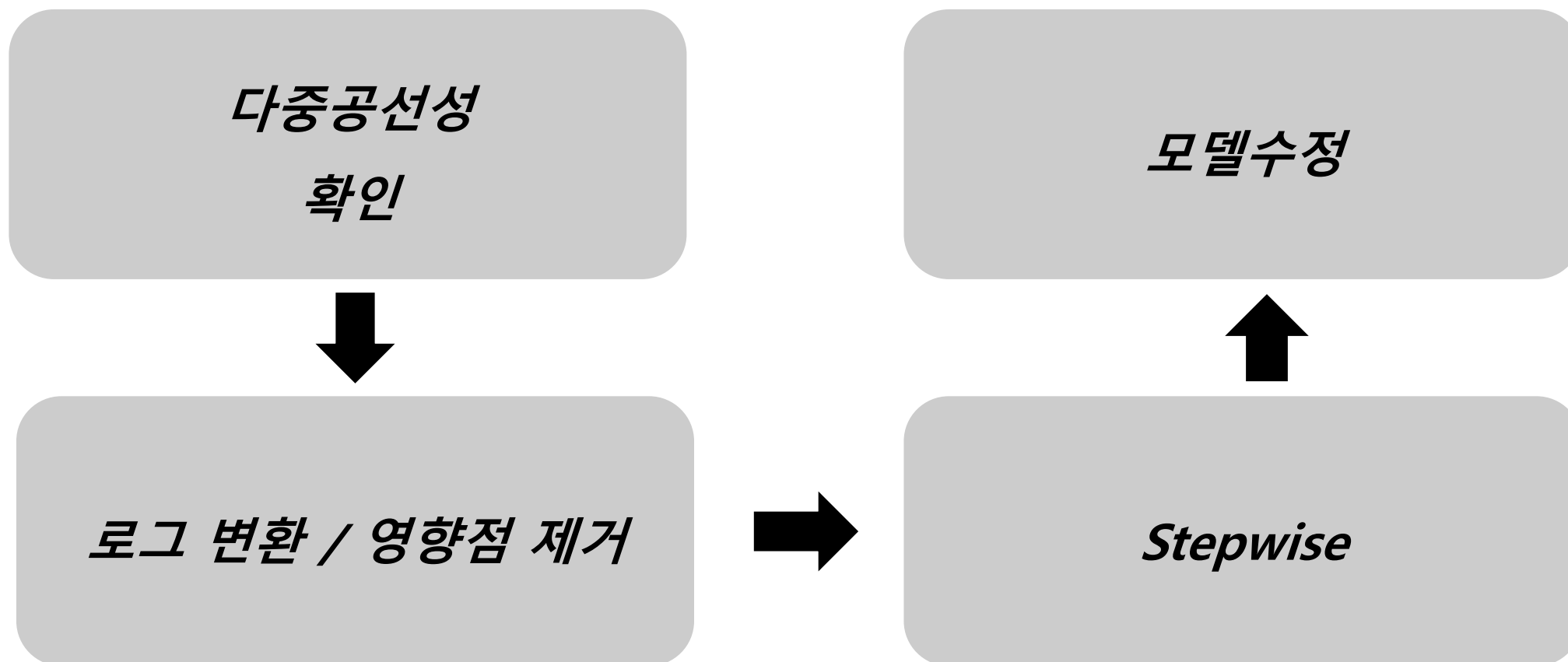
03 **데이터 분석**

04 해석



**STEPWISE 이후
만들어진 모델이
적합한지
확인하는 과정**







다중공선성 문제

VARIABLE	GVIF	Df	GVIF ^{1/(2*Df)}
...			
RoofStyle	1.212069	1	1.100940
MasVnrArea	1.445779	1	1.202406
BsmtQual	9.533468	3	1.456158
BsmtExposure	53.410344	3	1.940607
BsmtFinType1	64.346284	3	2.832245
TotalBsmtSF	3.603003	1	1.898158
OverallCond	1.793488	2	1.157243
...			

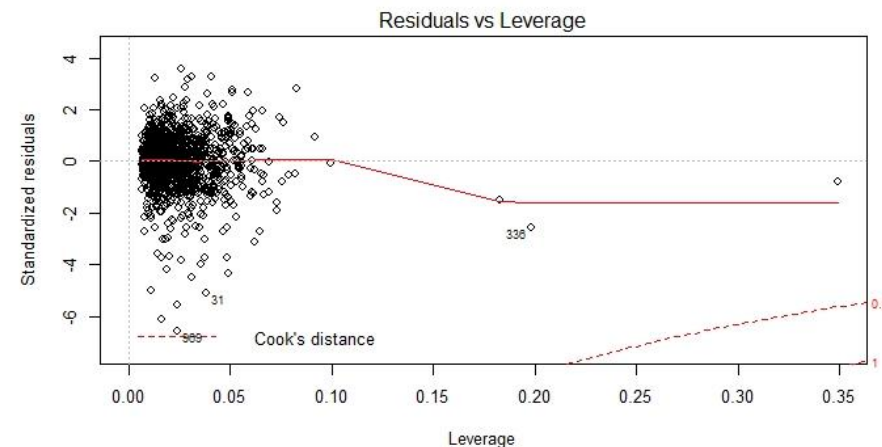
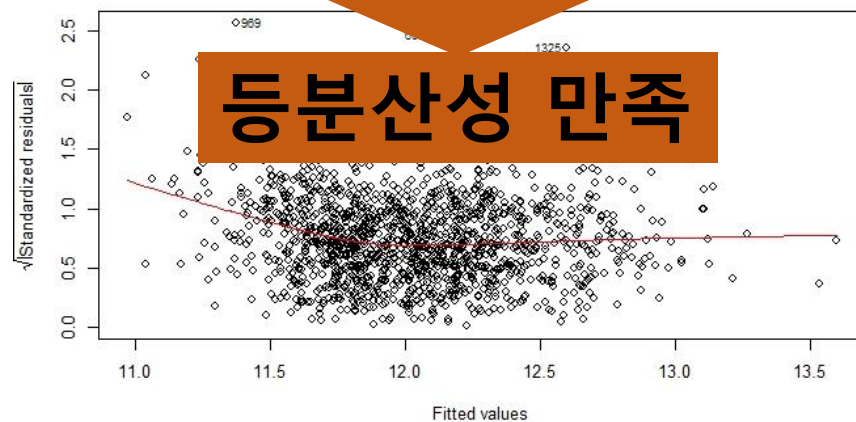
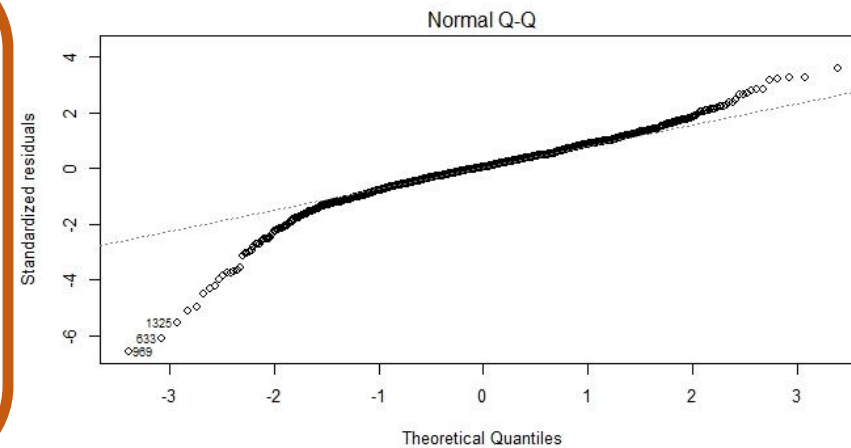
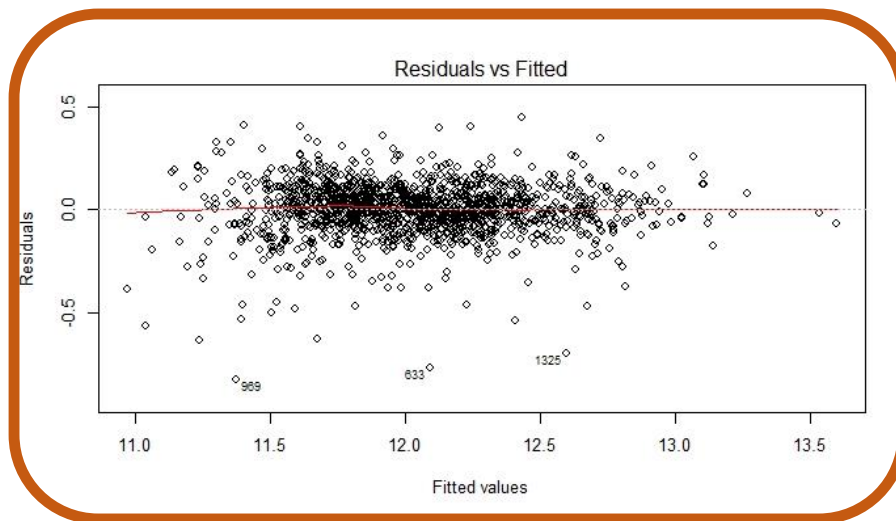
변수제거



로그 변환

영향점 제거

➡ 잔차그림





ANOVA 분석을 통한 모델 수정

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-1.620e+05	1.185e+05	-1.366	0.172068	
LotArea	2.820e-01	9.193e-02	3.067	0.002210	*
LotShape1	2.013e+03	1.830e+03	1.100	0.271456	
LotShape2	1.326e+04	5.109e+03	2.596	0.009532	*
Neighborhood1	1.211e+04	2.567e+03	4.720	2.60e-06	**
Neighborhood2	2.265e+04	3.017e+03	7.506	1.07e-13	**
Neighborhood3	4.026e+04	3.883e+03	10.369	< 2e-16	**
Condition11	-5.768e+03	5.606e+03	-1.029	0.303715	
Condition12	8.608e+03	4.719e+03	1.824	0.068384	

P-value가 높은 변수들에 대한
anova분석 실시

Anova검정 결과
P-value가 0.05보다
높은 변수 제거



최종 회귀모델

	Estimate
(Intercept)	10.92457
OverallQual3	0.2270938
OverallCond1	0.1873151
Neighborhood3	0.1778654
OverallQual2	0.1487568
...	...
GrLivArea	0.0000240
TotalBsmntSF	0.0000167
Lotarea	0.0000001
...	...

Contents



01 분석 목적 및 데이터 소개

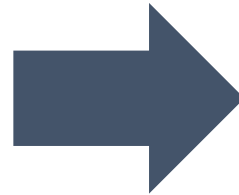
02 데이터 전처리

03 데이터 분석

04 해석



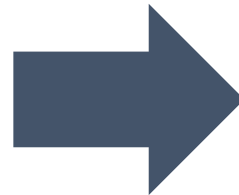
집의 재료, 완성도
집의 상태
위치 지역



집값에 가장
큰 영향








총 생활면적
땅(부지) 면적
지하의 총 면적



영향 **X**



kaggle 에 적용

2359	▼ 292	visheshwar		0.14040	7	2mo
2360	▼ 292	Amardeep Singh		0.14040	6	1mo
2361	new	EUNTAE		0.14043	3	~10s
Your Best Entry ↑ Your submission scored 0.14067, which is not an improvement of your best score. Keep trying!						
2362	▼ 293	qilowa		0.14045	3	18d
2363	new	Dillon Gash		0.14045	3	10d



감사합니다