HOUSE PRICES

강병욱, 김기훈, 김수정, 김태현, 임규리

INDEX

1. 데이터 전처리
    1. NA 처리
    2. Categoric to Numeric
    3. Categoric to Numeric – Adding Custom Numeric Features
    4. Treating Outliers

2. Preprocessing
    1. Kolmogorov-Smirnov Test

KERNELS

Tanner Carbonati

Detailed Data Analysis & Ensemble Modeling

last run 9 months ago · R notebook · 25676 views
using data from House Prices: Advanced Regression Techniques · Public

Stephanie Kirmer

Fun with Real Estate Data

last run 2 years ago · R notebook · 29271 views
using data from House Prices: Advanced Regression Techniques · Public

Erik Bruin

House prices: Lasso, XGBoost, and a detailed EDA

last run a day ago · R notebook · 13112 views
using data from House Prices: Advanced Regression Techniques · Public

1. 데이터 설명

81 variables

Target Variable

# of data

2919

ID	MSSubClass	MSZoning	LotFrontage	•••	SaleCondition	SalePrice
1	60	RL				208,500
2	20	RL				181,500
3	60	RL				223,500
⋮	⋮	RM				...
2919	45	RL				118000

Train set

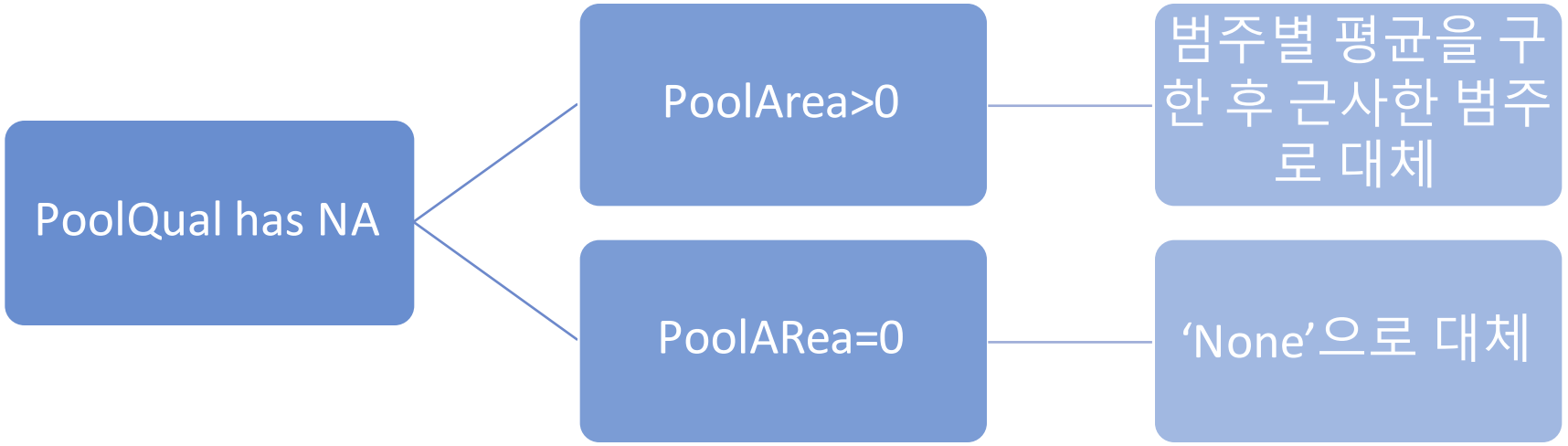
- Row - 1460
- Col - 81

Test set

- Row - 1459
- Col - 80

1. 데이터 전처리

1. NA(결측치) 처리 – 다른 변수와의 관계를 이용




```
# A tibble: 4 x 3
  PoolQC     mean counts
  <chr>     <dbl>  <int>
1 Ex         360.      4
2 Fa         584.      2
3 Gd         648.      4
      PoolQC PoolArea
2421    <NA>      368
2504    <NA>      444
2600    <NA>      561
```

```
df.combined[2421,'PoolQC'] = 'Ex'
df.combined[2504,'PoolQC'] = 'Ex'
df.combined[2600,'PoolQC'] = 'Fa'
```

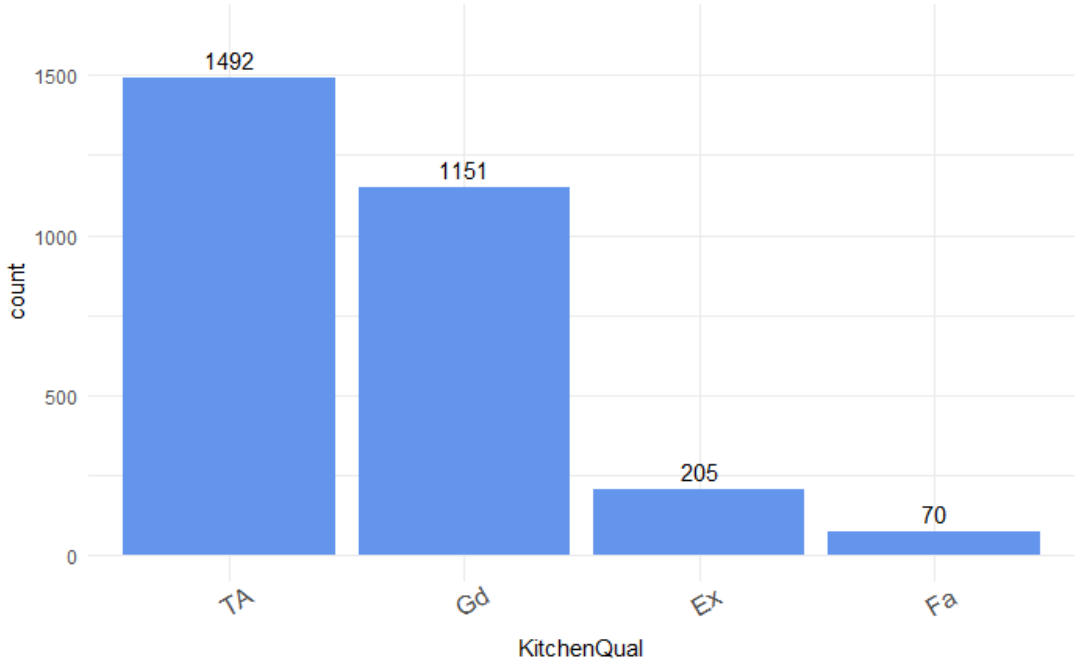
```
df.combined$PoolQC[is.na(df.combined$PoolQC)] = 'None'
```

1. 데이터 전처리

1. NA(결측치) 처리 – 최빈값으로 대체

1. KitchenQual, Electrical 등에서 1개의 NA

   →최빈값으로 대체




```
```{r warning=FALSE, message=FALSE}
plot.categoric('KitchenQual', df.combined)
df.combined$KitchenQual[is.na(df.combined$KitchenQual)] = 'TA'

plot.categoric('Electrical', df.combined)
df.combined$Electrical[is.na(df.combined$Electrical)] = 'SBrkr'
```
```

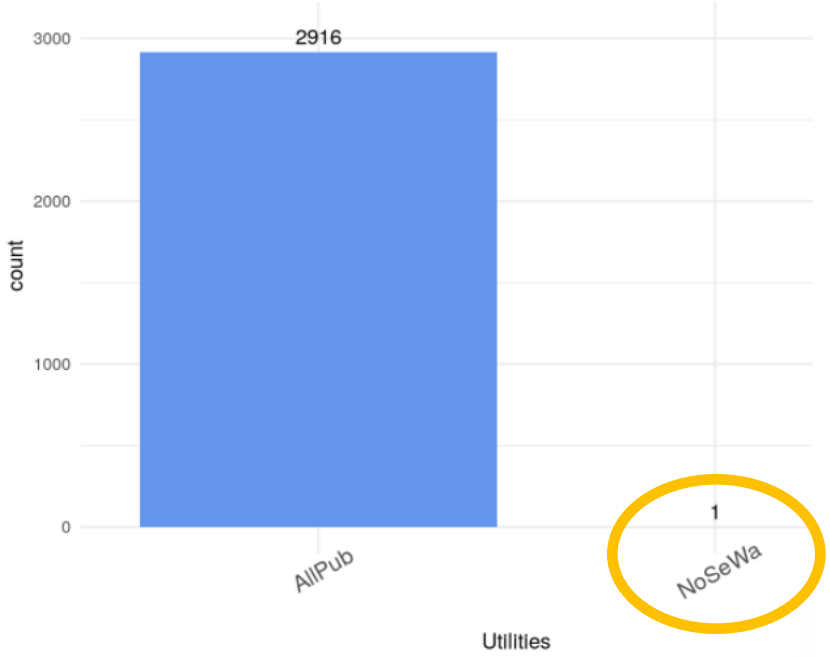
1. 데이터 전처리

1. NA(결측치) 처리 – 임의처리 및 변수삭제

- Exterior1st, 2nd는 한 집의 값만 비어 있으며, 이 변수는 유용한 변수로 판단되는 수준이 없기 때문에 NA를 “Other”로 처리함
- Utilities는 train셋에 1개의 NeSeWa값이 존재, test셋은 모두 AllPub이므로 변수의 의미가 없음 => Utilities 변수 제거

Exterior NA’s

	Exterior1st	Exterior2nd
2152	NA	NA



1. 데이터 전처리

1. NA(결측치) 처리 – 다른 변수와 관계 + 최빈값/중앙값

• SaleType 결측치 처리
- SaleCondition 과의 관계를 통해 처리(관련 변수)

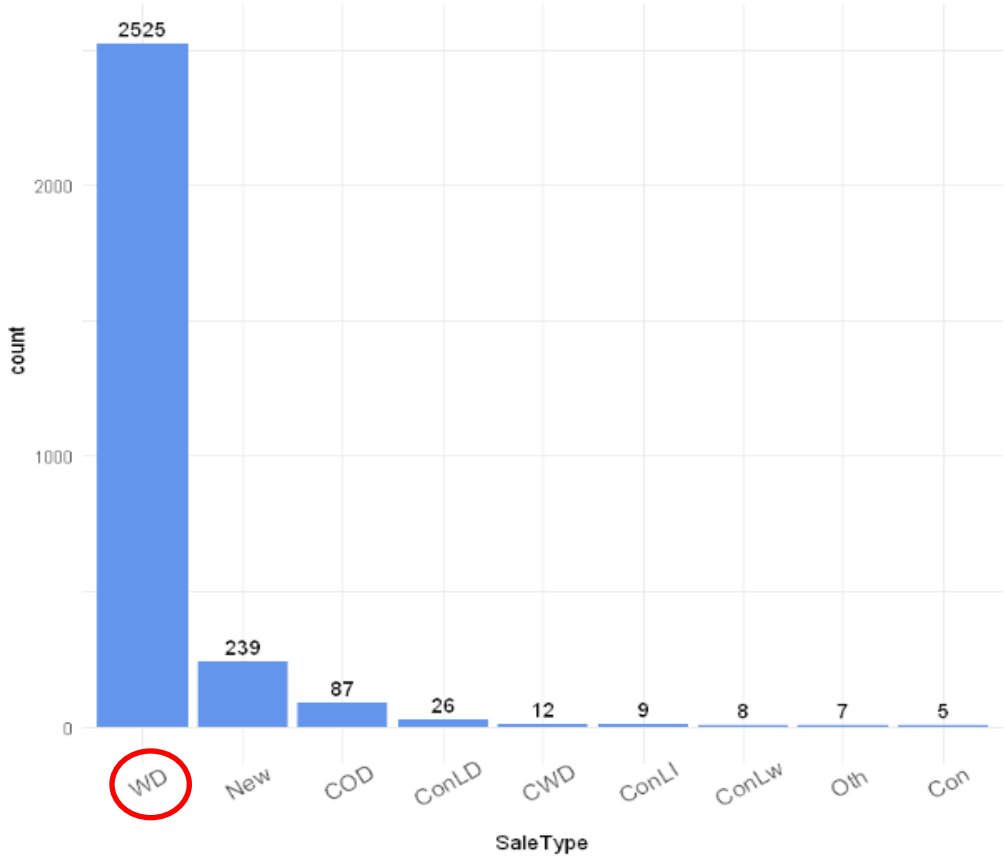
① SaleType의 결측값이 있는 SaleCondition

ID	SaleType	SaleCondition
2490	NA	Nominal

② SaleType 과 SaleCondition 의 관계 테이블 ➔ 최빈값 처리

	COD	Con	ConLD	ConLI	ConLw	CWD	New	Oth	**WD**
Abnorml	46	0	3	2	0	1	0	5	133
AdjLand	0	0	0	0	0	0	0	0	12
Alloca	0	0	0	0	0	0	0	0	24
Family	2	0	1	2	1	1	0	1	38
Normal	39	4	21	5	7	10	0	1	**2314**
Partial	0	1	1	0	0	0	239	0	4

③ SaleType 빈도수 plot



1. 데이터 전처리

1. NA(결측치) 처리 – 결측치 NONE 혹은 0 으로 처리

MasVnrType & MasVnrArea

na.omit(MasVnrType) <chr>	MedianArea <dbl>	counts <int>
None	0	1742
BrkCmn	161	25
Stone	200	249
BrkFace	203	879

4 rows

```
```{r warning=FALSE, message=FALSE}
df.combined$MasVnrType[is.na(df.combined$MasVnrType)] = 'None'
df.combined$MasVnrArea[is.na(df.combined$MasVnrArea)] = 0
```
```

변수 24개 중 하나를 제외하고 두 변수 모두 NA인 경우이므로, 해당사항이 없는 경우라고 가정하고, 면적이 198인 관측치에 대해 타입 별 중앙값을 구해 가까운 값으로 대체함

1. 데이터 전처리

2. CATEGORIC TO NUMERIC

1. 데이터 분리

- 변수에 값이 없는 전처리 이후 데이터를 numeric set과 categoric set 으로 나눔
- One-hot encoding을 이용하여 categoric set을 2진수로 변경.
- 새로운 데이터 프레임에 numeric 데이터를 모두 합산하여 예측 모델 적용 시 활용.
- Numeric 으로 새로운 데이터 프레임 생성

```
num_features <- names(which(sapply(df.combined, is.numeric)))
cat_features <- names(which(sapply(df.combined, is.character)))

df.numeric <- df.combined[num_features]
```

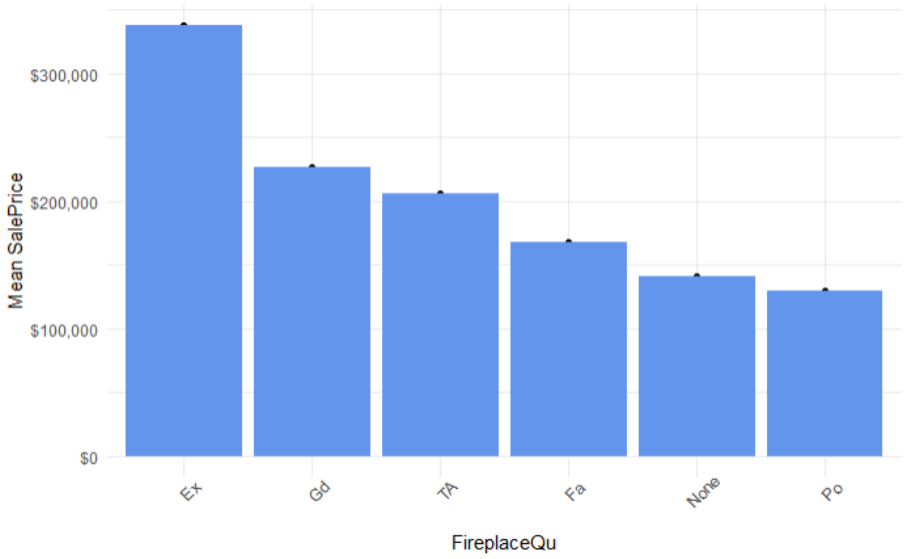
1. 데이터 전처리

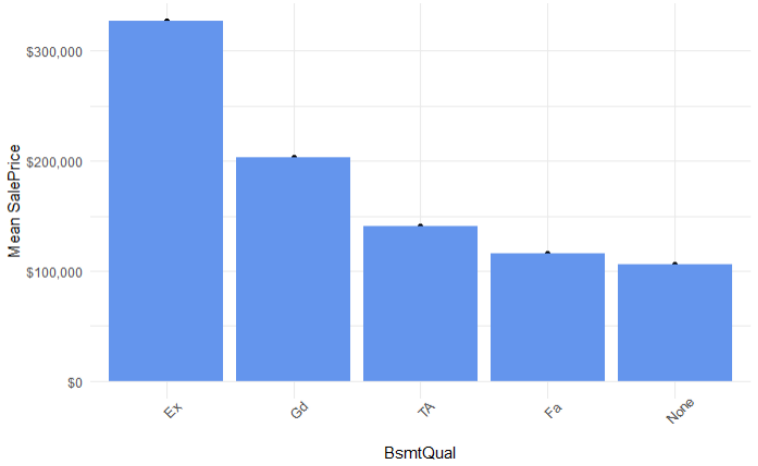
2. CATEGORIC TO NUMERIC

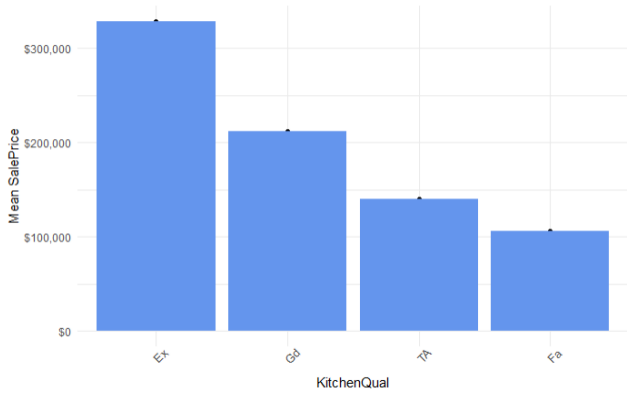
2. 등급으로 측정 가능한 데이터

- 등급으로 측정이 가능한 데이터들은 numeric 으로 변경.
- 변수 카테고리들을 1,2,3...n으로 정렬하여 예측해야 하는 SalePrice의 중앙 값과 OverallQual의 평균값으로 구함.
- 값이 높을수록 높은 점수를 부여

- (None' = 0, 'Po' = 1, 'Fa' = 2, 'TA' = 3, 'Gd' = 4, 'Ex' = 5)





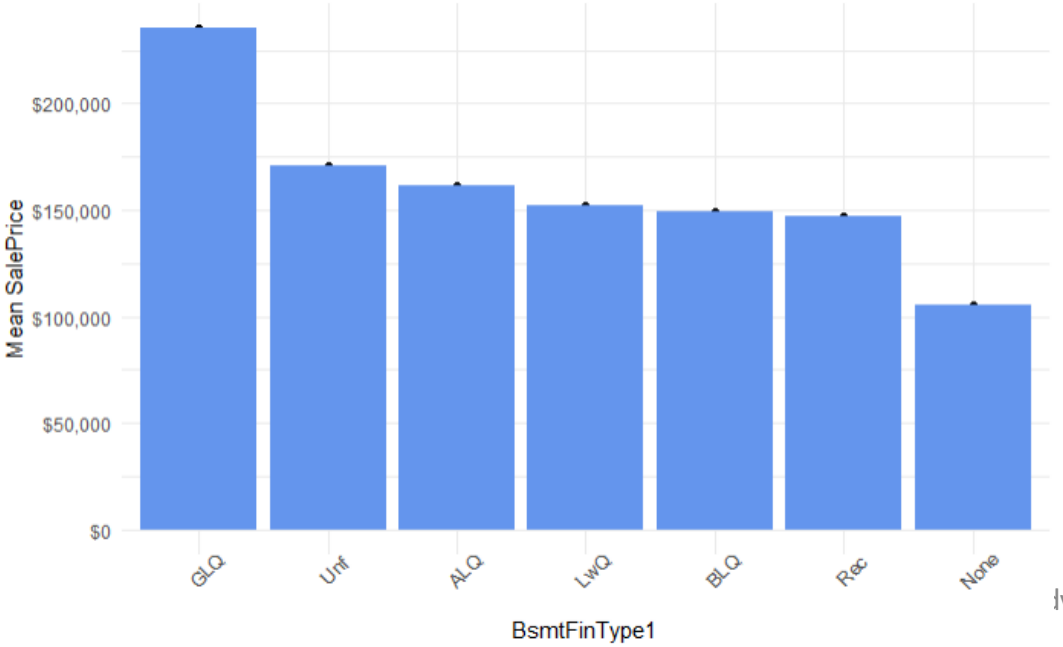


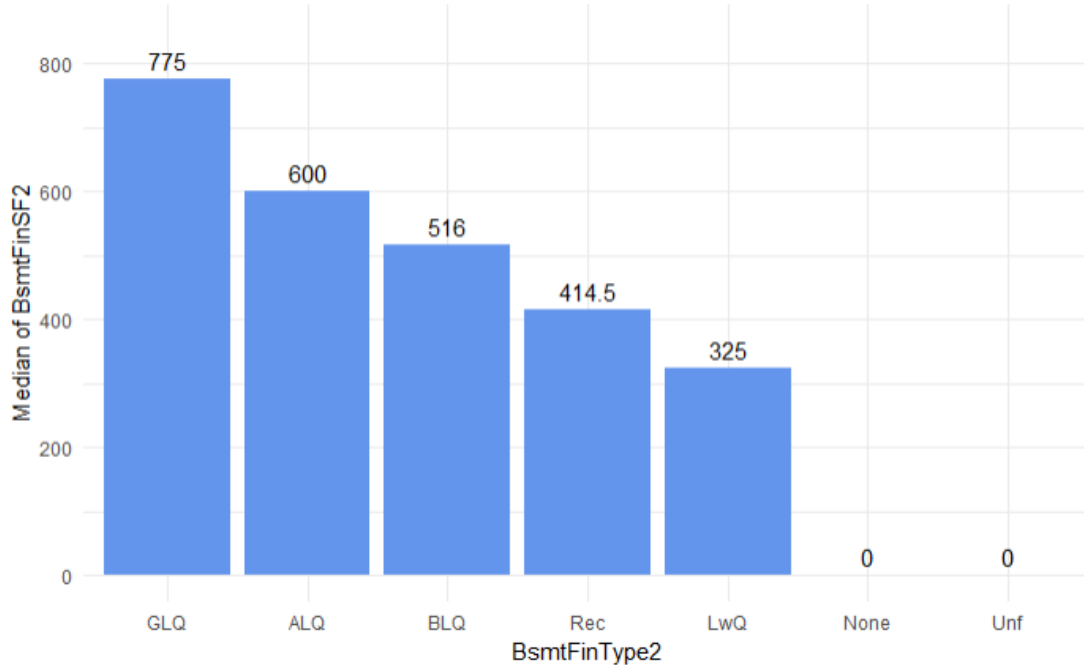
1. 데이터 전처리

2. CATEGORIC TO NUMERIC

2. 등급으로 측정 가능한 데이터

- 변수 `BsmtFinType1` 과 `BsmtFineType2`는 1층,2층의 quality를 나타냄. 유사한 Quality의 따른 패턴이 있을 것으로 추정됨.
- Basement quality별 순서 *'None' < 'Unf' < 'LwQ' < 'BLQ' < 'Rec' < 'ALQ' < 'GLQ'*.
- bsmt.fin.list <- c('None' = 0, 'Unf' = 1, 'LwQ' = 2,'Rec'= 3, 'BLQ' = 4, 'ALQ' = 5, 'GLQ' = 6)





1. 데이터 전처리

2. CATEGORIC TO NUMERIC

2. 등급으로 측정 가능한 데이터

QUALITY 관련 변수 처리	
변수	처리
ExterQual	'None' =0 ,'Po' = 1, 'Fa' = 2, 'TA' = 3, 'Gd' = 4, 'Ex' = 5 (변수들이 위의 수준들을 항상 가짐)
ExterCond	
GarageQual	
GarageCond	
FirePlaceQu	
KitchenQual	
HeatingQC	
BsmtQual	

1. 데이터 전처리

2. CATEGORIC TO NUMERIC

2. 등급으로 측정 가능한 데이터

나머지 변수 처리	
변수	처리
BsmtExposure	'None' = 0, 'No' = 1, 'Mn' = 2, 'Av' = 3, 'Gd' = 4
BsmtFinType1	'None' = 0, 'Unf' = 1, 'LwQ' = 2,'Rec'= 3, 'BLQ' = 4, 'ALQ' = 5, 'GLQ' = 6
BsmtFinType2	'None' = 0, 'Unf' = 1, 'LwQ' = 2,'Rec'= 3, 'BLQ' = 4, 'ALQ' = 5, 'GLQ' = 6
Functional	'None' = 0, 'Sal' = 1, 'Sev' = 2, 'Maj2' = 3, 'Maj1' = 4, 'Mod' = 5, 'Min2' = 6, 'Min1' = 7, 'Typ'= 8
GarageFinish	'None' = 0,'Unf' = 1, 'RFn' = 1, 'Fin' = 2
Fence	'None' = 0, 'MnWw' = 1, 'GdWo' = 1, 'MnPrv' = 2, 'GdPrv' = 4
NewerDwelling	'20‘, '60‘, 120' = 1, 나머지 0

1. 데이터 전처리

3. CATEGORIC TO NUMERIC – ADDING CUSTOM NUMERIC FEATURES

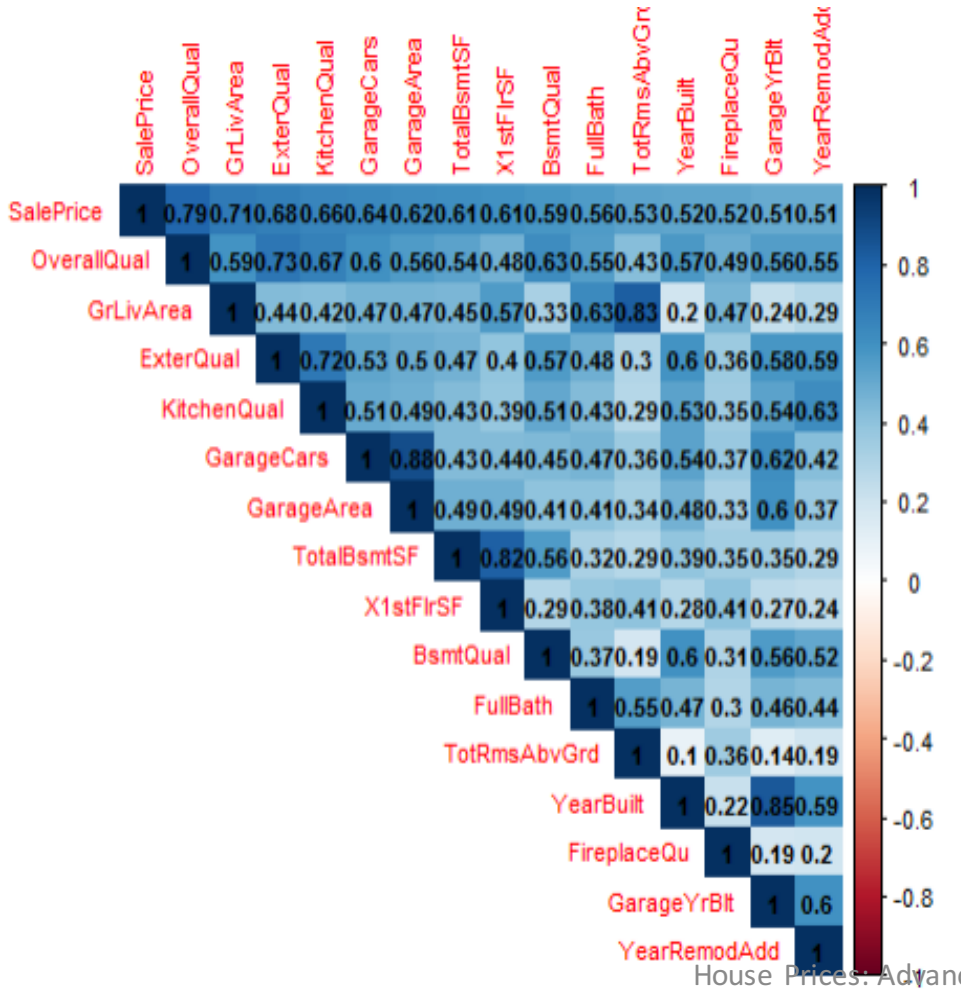
1. SALEPRICE와의 상관관계 TOP 10 분석

- 데이터를 numeric으로 변경 후 SalePrice와의 높은 상관관계 변수를 확인
- 상관관계 확인을 위해 2개의 변수의 상관계수를 비교
- 두개의 변수의 상관계수가 0인 경우 선형 관계가 없다고 해석
- 상관 계수가 0과 1사이면 positive linear relationship, 0과 -1사이면 negative linear relationship 으로 해석
- SalePrice 변수와의 상관 계수가 0.5보다 크거나 -0.5보다 작은 변수를 찾아야함.

1. 데이터 전처리

3. CATEGORIC TO NUMERIC – ADDING CUSTOM NUMERIC FEATURES

1. SALEPRICE와의 상관관계 TOP 10 분석

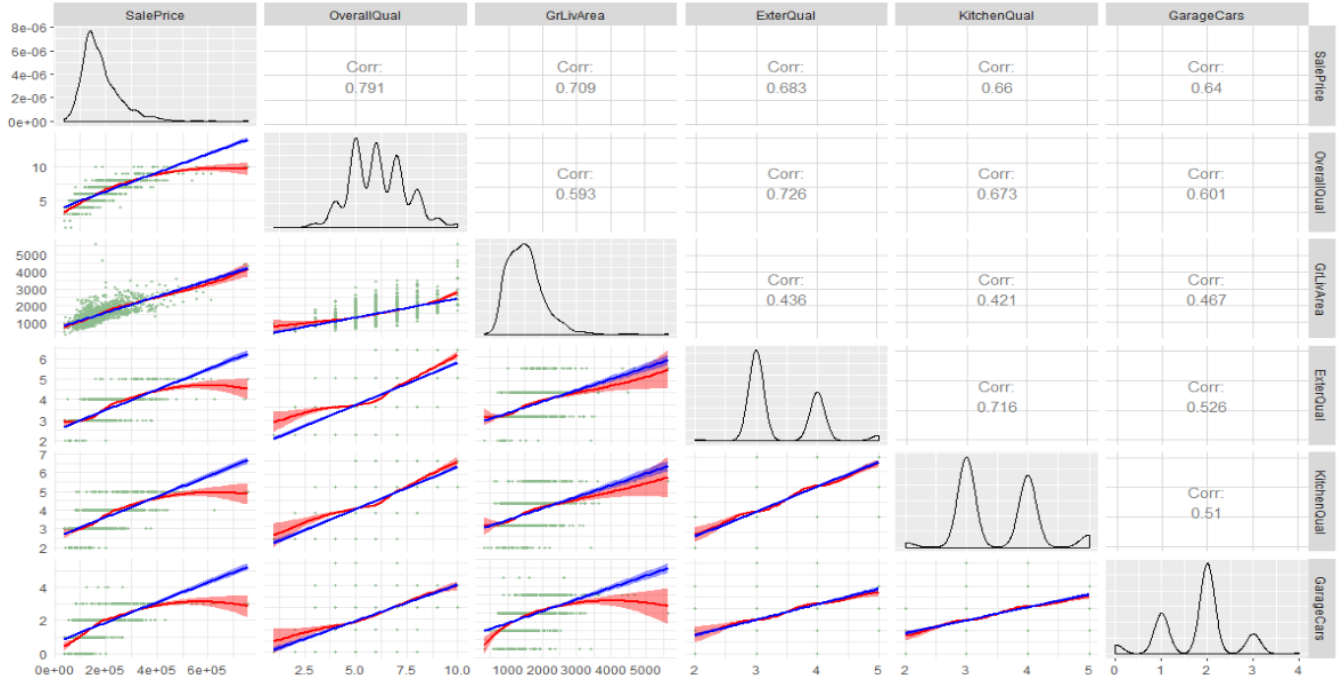


- OverallQual & GrLivArea: 상관 계수가 가장 높음
- GarageCars & GarageArea: 주차장의 차와 주차장 넓이 변수 끼리의 상관계수도 높음
- TotalBsmtSF & 1stFlrSF: 집의 면적이 집 가격과의 관계가 있는 것으로 보임. Basement 또한 집의 1층 면적과 관계가 있기에 두개의 변수 끼리도 상관 계수가 높음
- FullBath & TotRmsAbvGrd: 집의 방 개수 또한 집 가격과 상관 계수가 높음. GrLivArea(집 면적) 변수와 TotRmsAbvGrd(방 개수) 변수 끼리 상관계수가 높음. 방이 많으나 집 면적이 적을 경우 집 가격에 어떠한 영향이 있는지 확인하는 것도 의미 있을 듯
- YearBuilt & YearRemodAdd: 집 연도가 현재 시점과 가까울 수록 집 가격이 높은 것으로 보임.

1. 데이터 전처리

3. CATEGORIC TO NUMERIC – ADDING CUSTOM NUMERIC FEATURES

1. SALEPRICE와의 상관관계 TOP 10 분석

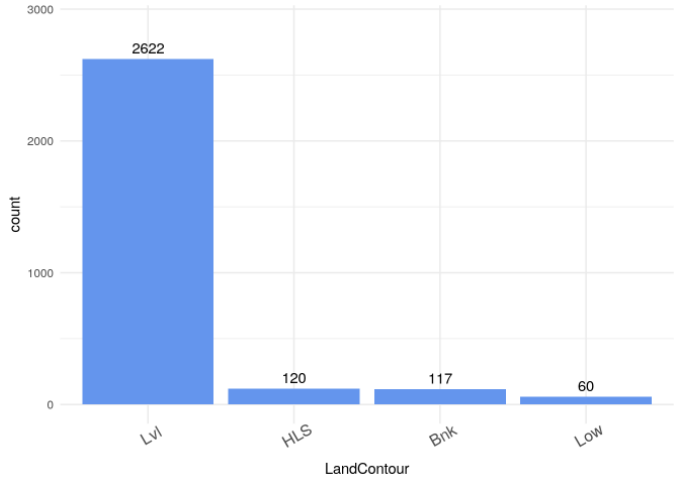


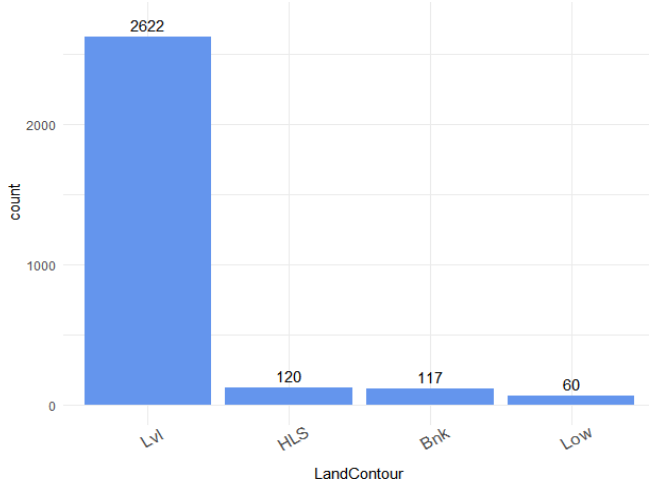
- 그래프의 파란색 선은 단순 선형 회귀를 나타내며, 빨간 선은 다항 회귀를 나타냄.
- OverallQual, GrLivArea 변수에 선형 모델 선을 볼 수 있으나 아웃라이어가 있어 들여다 볼 필요가 있음.
- 몇몇 집들 중 overall Quality가 10이지만, 집 가격이 특이하게 낮게 보임.
- GrLivArea, TotalBsmtSF, GarageCars, GarageArea 변수에도 이러한 현상이 보임.
- 집 면적이 넓거나 집의 주차 가능한 4개 차량의 주차장이 있어도 집 가격이 높아 보이지 않은 것으로 보임.

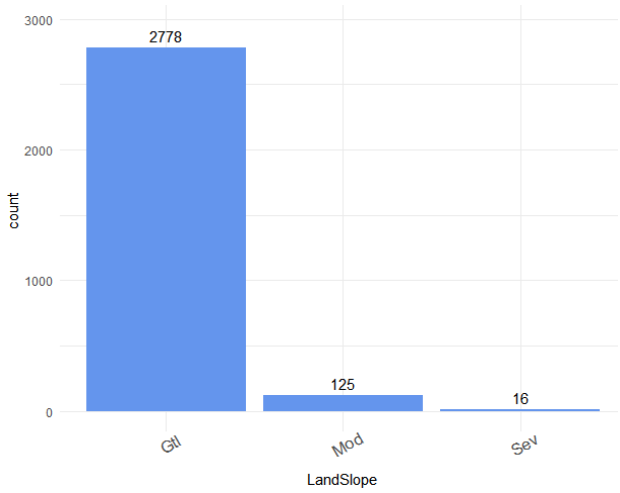
1. 데이터 전처리

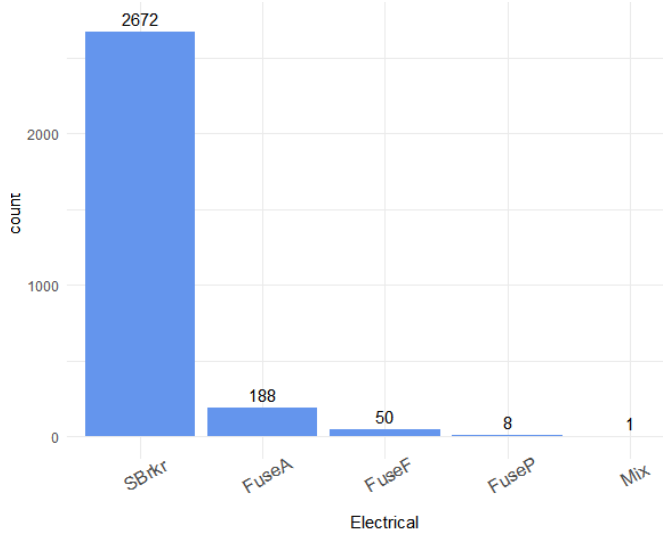
3. CATEGORIC TO NUMERIC – ADDING CUSTOM NUMERIC FEATURES

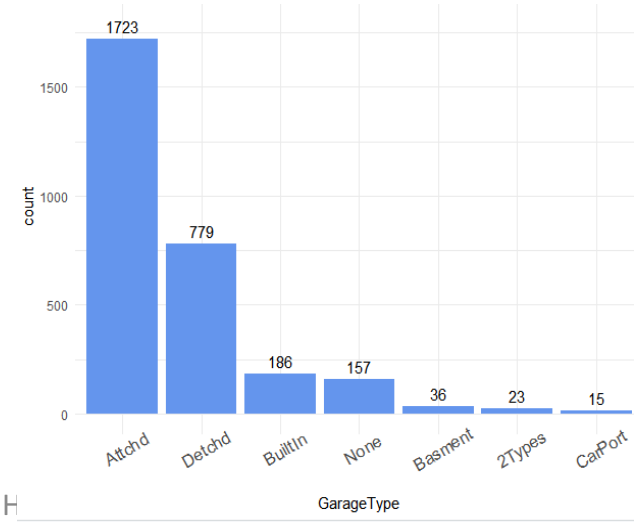
2. ONE HOT ENCODING FOR SOME NOMINAL VALUES

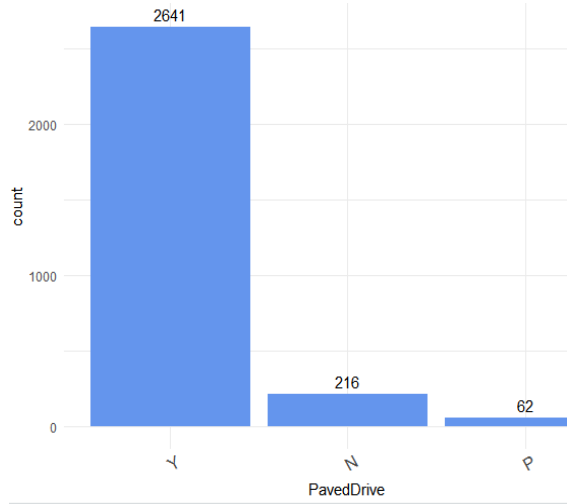












1. 데이터 전처리

3. CATEGORIC TO NUMERIC – ADDING CUSTOM NUMERIC FEATURES

2. ONE HOT ENCODING FOR SOME NOMINAL VALUES

변수	새로운 변수명	처리
LotShape	RegularLotShape	Reg = 1 , 나머지 0
LandContour	LandLeveled	Lvl= 1 , 나머지 0
LandSlope	LandSlopeGentle	Gtl= 1 , 나머지 0
Electrical	ElectricalSB	SBrkr = 1 , 나머지 0
GarageType	GarageDetchd	Detchd = 1 , 나머지 0
PavedDrive	HasPavedDrive	Y = 1 , 나머지 0
WoodDeckSF	HasWoodDeck	0이상 = 1, 나머지 0
2ndFlrSF	Has2ndFlrSF	0이상 = 1, 나머지 0
MasVnrArea	HasMasVnr	0이상 = 1, 나머지 0

1. 데이터 전처리

3. CATEGORIC TO NUMERIC – ADDING CUSTOM NUMERIC FEATURES

2. ONE HOT ENCODING FOR SOME NOMINAL VALUES

변수	새로운 변수명	처리
MiscFeature	HasShed	Shed = 1 , 나머지 0
YearBuilt != YearRemodAdd	Remodeled	1, 나머지 0
YearRemodAdd >= YrSold	RecentRemodel	1, 나머지 0
YearBuilt == YrSold	NewHouse	1, 나머지 0
SaleCondition	PartialPlan	Partial
HeatingQC	HeatingScale	'Po' = 0, 'Fa' = 1, 'TA' = 2, 'Gd' = 3, 'Ex' = 4

3. ADDING DUMMY VARIABLES

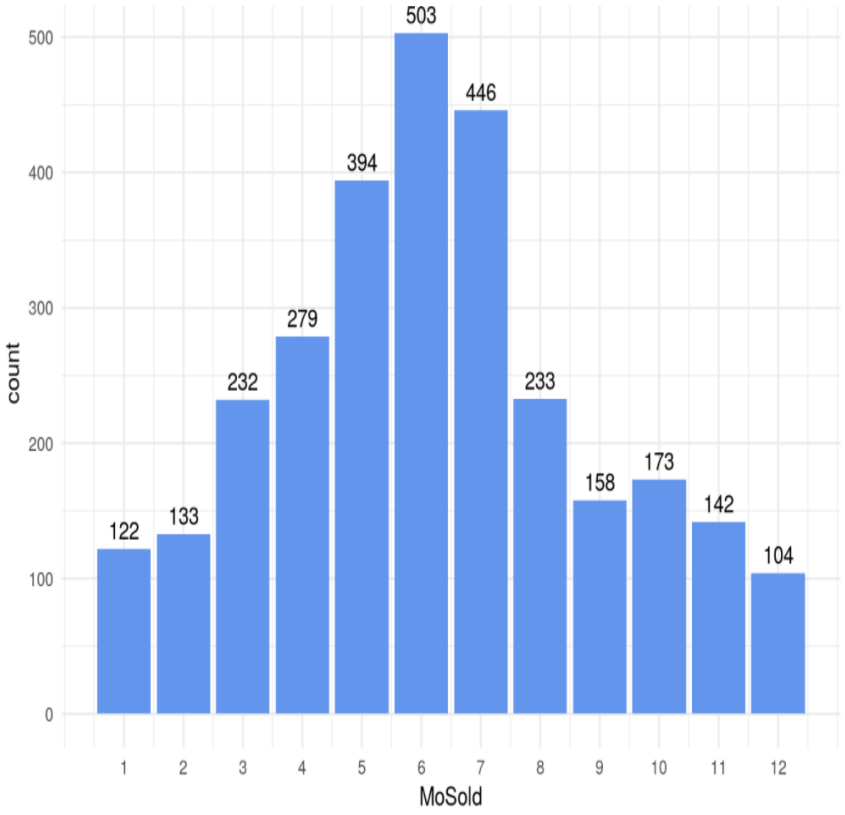
- 아래에 있는 면적과 관련된 변수는 값이 0이면 0, 0보다 크면 1이란 Dummy Variable을 생성
- 'X2ndFlrSF', 'MasVnrArea', 'WoodDeckSF', 'OpenPorchSF', 'EnclosedPorch', 'X3SsnPorch', 'ScreenPorch'
- 새로운 변수명 앞에 'Has'를 붙여 구분

1. 데이터 전처리

3. CATEGORIC TO NUMERIC – ADDING CUSTOM NUMERIC FEATURES

3. ADDING DUMMY VARIABLES

변수	새로운 변수명	처리
MoSold	HighSeason	5,6,7월 = 1., 나머지 0

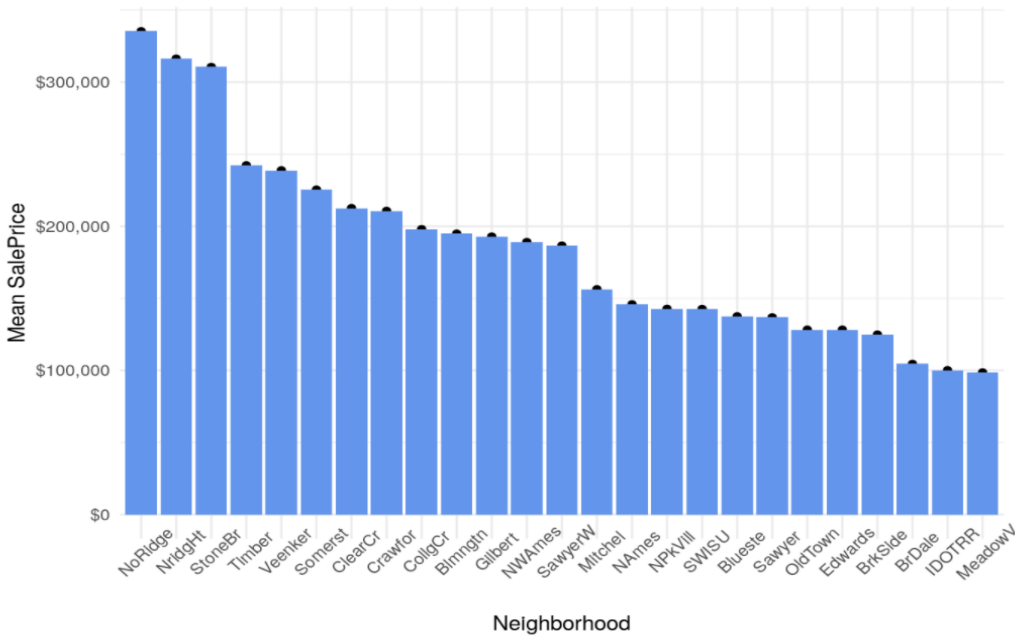


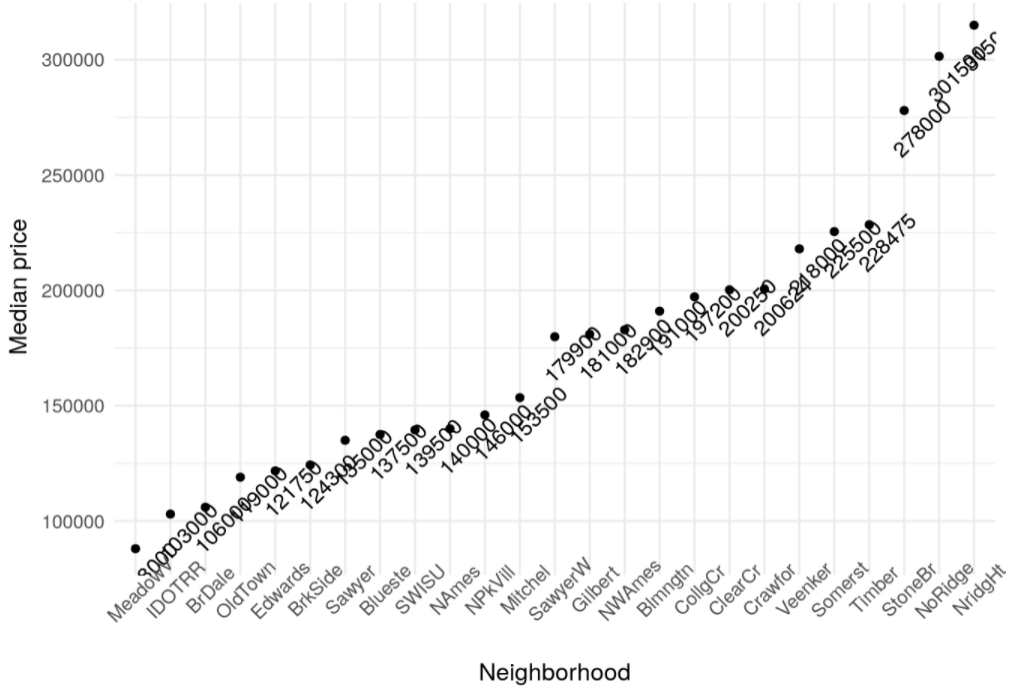
- 팔린 월의 경우 주로 여름철에 집값이 높음으로 5,6,7 월에 대한 더미 변수를 생성

1. 데이터 전처리

3. CATEGORIC TO NUMERIC – ADDING CUSTOM NUMERIC FEATURES

4. NEIGHBORHOOD





- 'Crawfor', 'Somerst, Timber', 'StoneBr', 'NoRidge', 'NridgeHt 의 값들을 rich 더미로 코딩
- Neighbor를 plot에 따라 0-4의 수치형 변수로 변경

1. 데이터 전처리

3. CATEGORIC TO NUMERIC – ADDING CUSTOM NUMERIC FEATURES

5. OTHER NEW NUMERICAL VARIABLES

- TotalArea : LotFrontage+LotArea+MasVnrArea+BsmtFinSF1+BsmtFinSF2+BsmtUnfSF+TotalBsmtSF+X1stFlrSF +X2ndFlrSF+GrLivArea+GarageArea+WoodDeckSF+OpenPorchSF+EnclosedPorch+X3SsnPorch+S creenPorch+LowQualFinSF+PoolArea

- AreaInside : 1stFlrSF + 2ndFlrSF

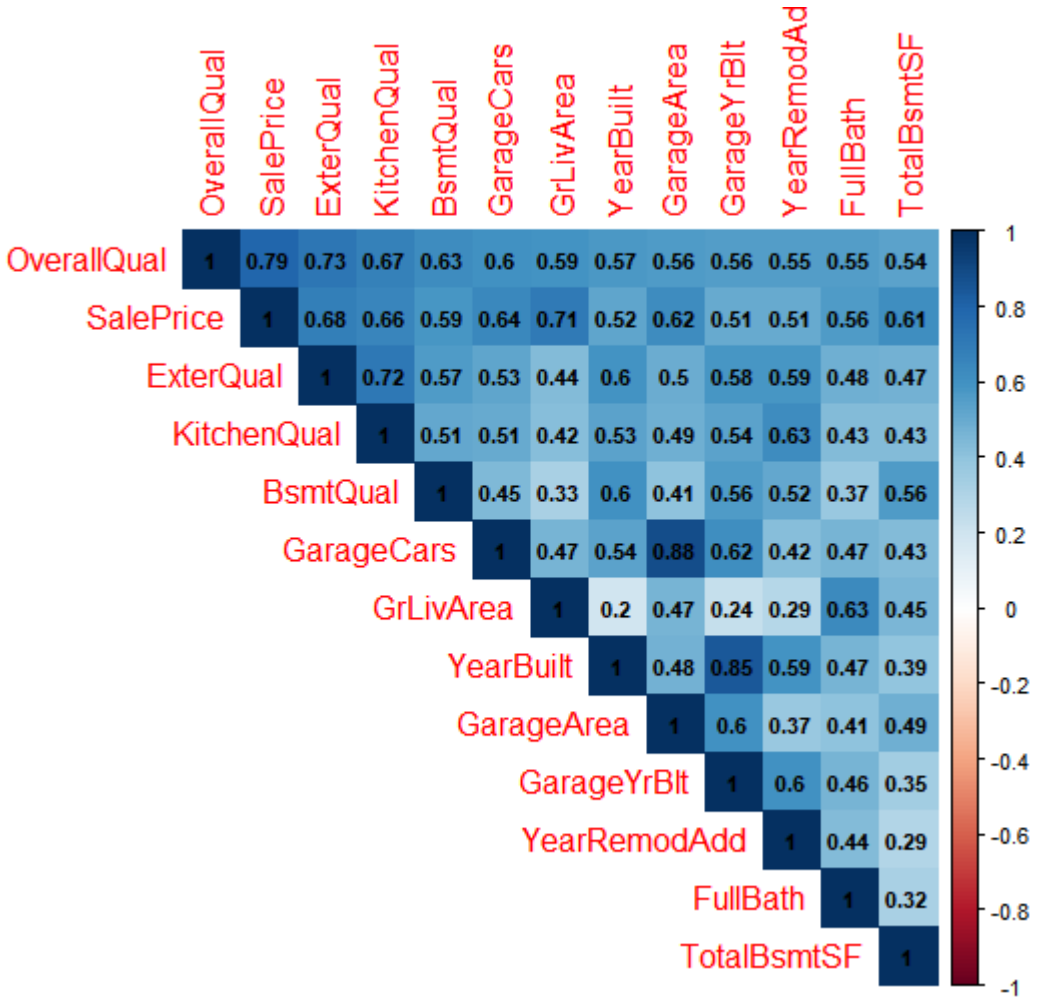
- Age : 2010 – YearBuilt
- (최근에 지어수록 SalePrice가 높다)

- TimeSinceSold : 2010 - YrSold

1. 데이터 전처리

3. CATEGORIC TO NUMERIC – ADDING CUSTOM NUMERIC FEATURES

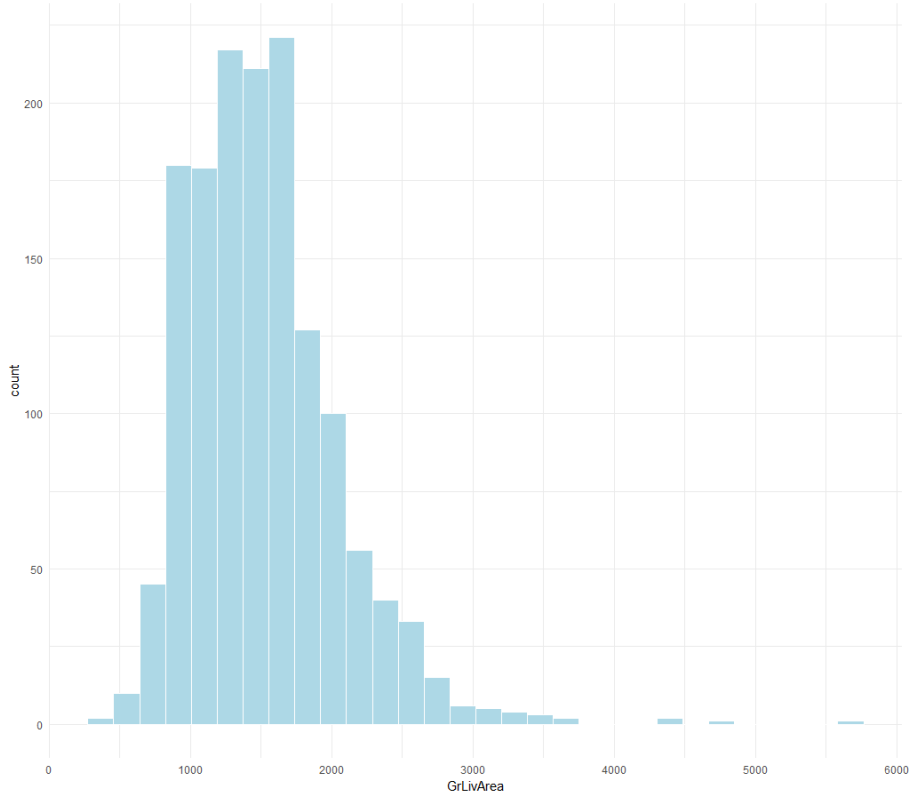
6. 새로운 상관관계 분석

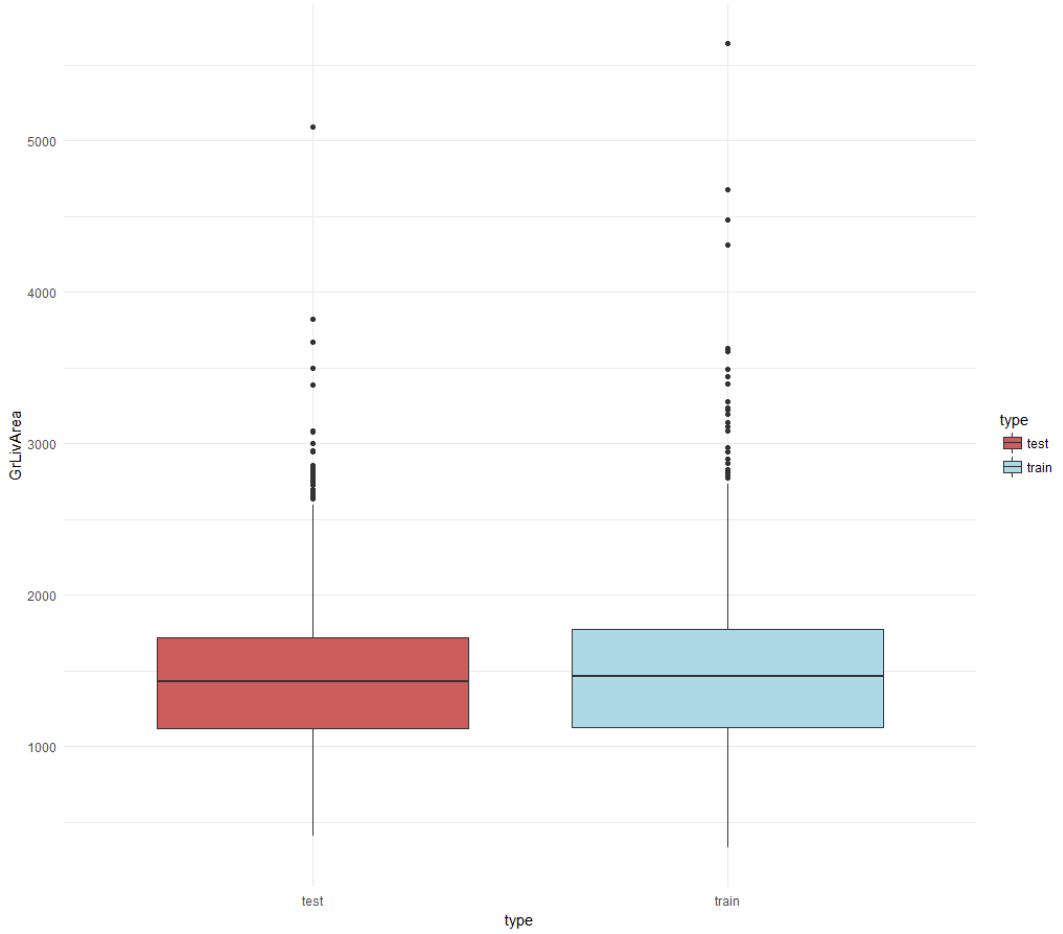


(새로운 상관계수)

1. 데이터 전처리

4. TREATING OUTLIERS

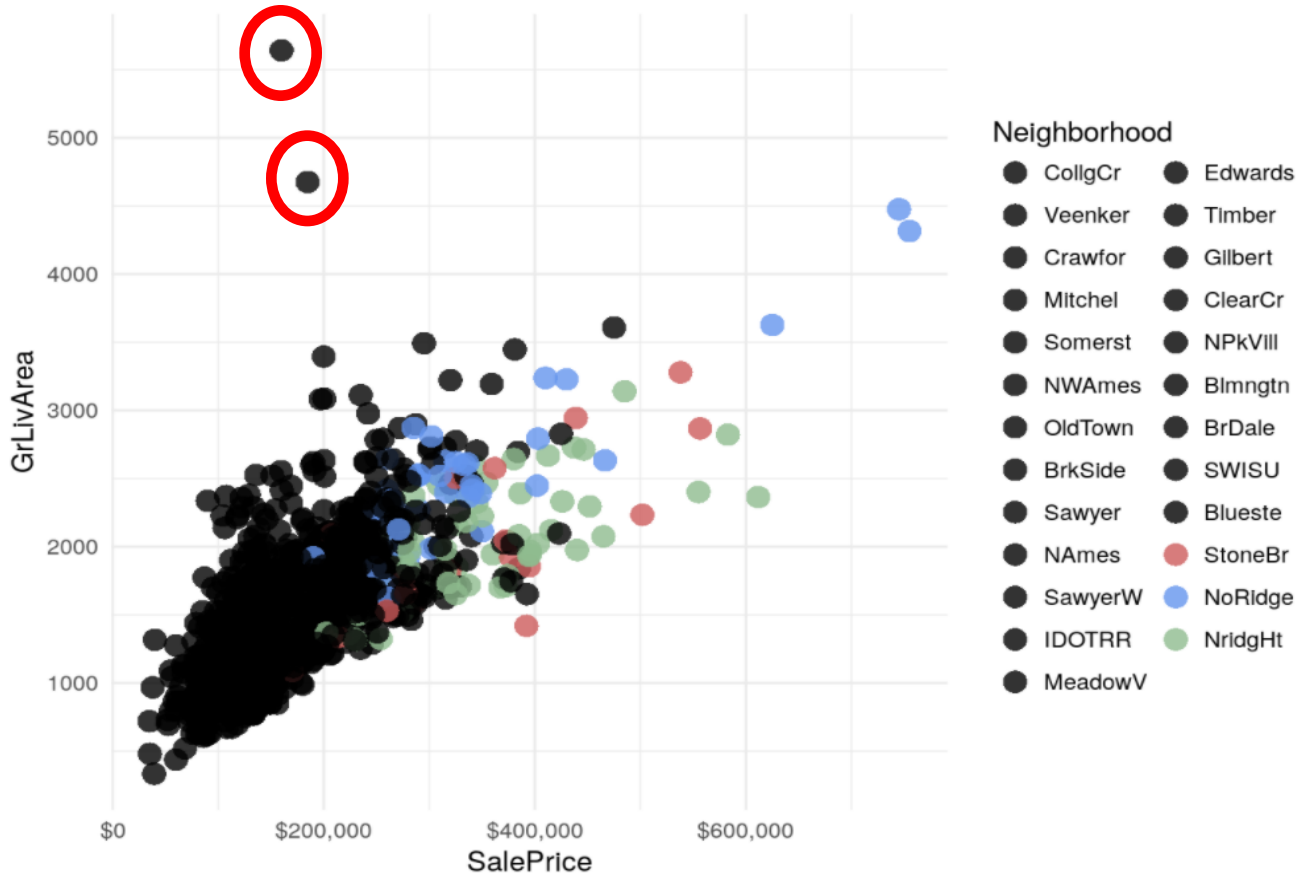




- Training에 GrLivArea > 4000인 데이터 4개, test에 한 개.
- SalePrice and GrLivArea에 skewness 를 야기

1. 데이터 전처리

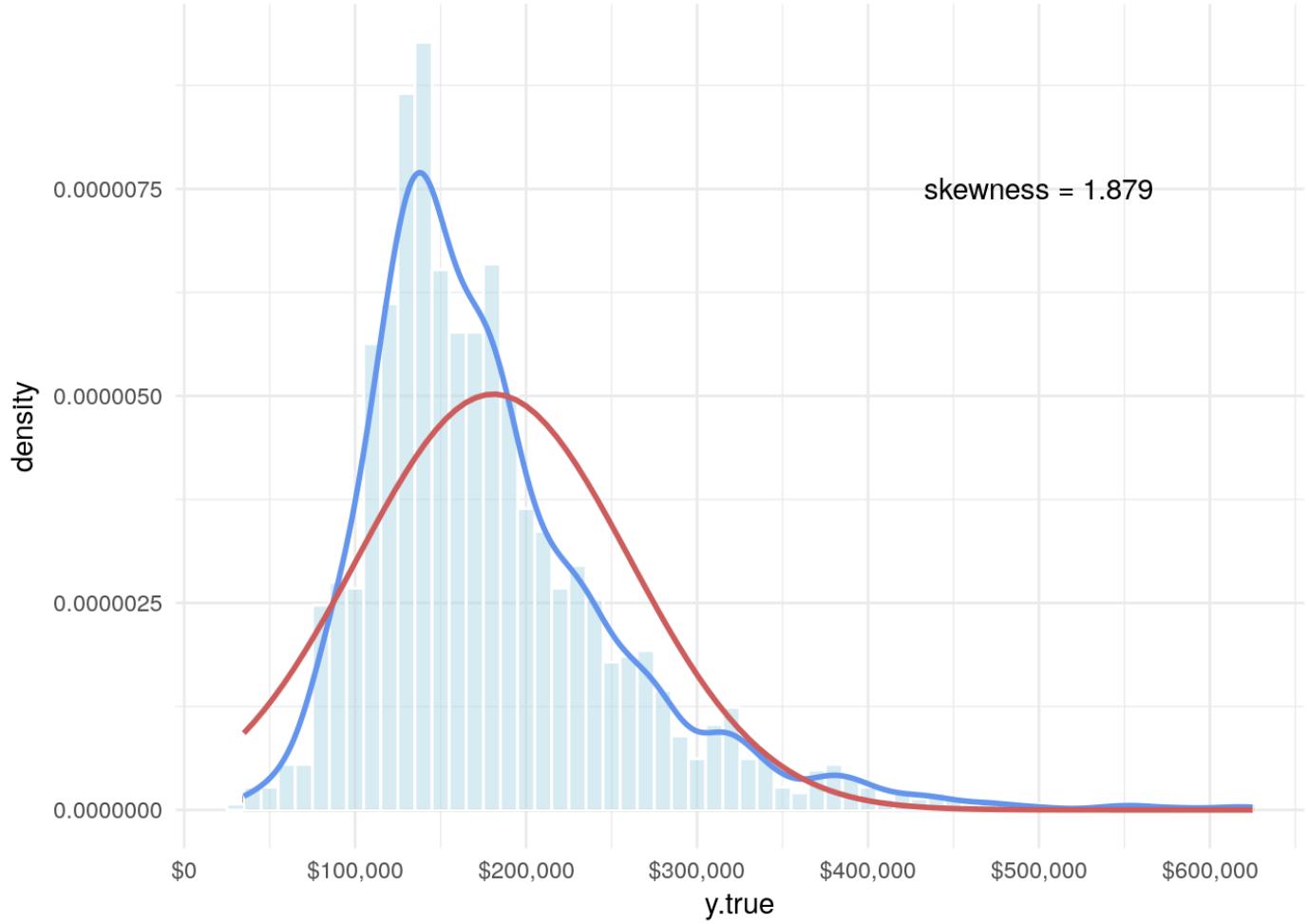
4. TREATING OUTLIERS



- 이들 중 특히 GrLivArea > 4000 &집값이 낮은 두 관측치는 두 변수의 상관관계에 제한
- 아웃라이어를 제거

2. PREPROCESSING

1. 콜모고로프-스미노프 검정(KOLMOGOROV-SMIRNOV TEST)



감사합니다

질문은..또르륵..