# 1주차 과제

김유진

21.08.01

# EDA (1) 데이터 구조 파악

| | Id | MSSubClass | MSZoning | LotFrontage | LotArea | Street | Alley | LotShape | LandContour |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 60 | RL | 65.0 | 8450 | Pave | NaN | Reg | Lvl |
| **1** | 2 | 20 | RL | 80.0 | 9600 | Pave | NaN | Reg | Lvl |
| **2** | 3 | 60 | RL | 68.0 | 11250 | Pave | NaN | IR1 | Lvl |
| **3** | 4 | 70 | RL | 60.0 | 9550 | Pave | NaN | IR1 | Lvl |
| **4** | 5 | 60 | RL | 84.0 | 14260 | Pave | NaN | IR1 | Lvl |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **1455** | 1456 | 60 | RL | 62.0 | 7917 | Pave | NaN | Reg | Lvl |
| **1456** | 1457 | 20 | RL | 85.0 | 13175 | Pave | NaN | Reg | Lvl |
| **1457** | 1458 | 70 | RL | 66.0 | 9042 | Pave | NaN | Reg | Lvl |
| **1458** | 1459 | 20 | RL | 68.0 | 9717 | Pave | NaN | Reg | Lvl |
| **1459** | 1460 | 20 | RL | 75.0 | 9937 | Pave | NaN | Reg | Lvl |

1460 rows × 81 columns

▶총 1460개의 행, 81개의 열로 구성

# EDA (1) 데이터 구조 파악

## Column들의 의미를 파악하기 위해, 변수들을 categorizing

- 지리적 정보: MSZoning(농업, 상업, 주거 밀도 정도), Neighborhood, LotFrontage, Street, Alley, LotConfig, Condition1, Condition2(근접 도로)

- 건축 형태 정보:
  - 건축물 형태: MSSubClass(건축 년도), LotShape, LandContour(평탄함 정도), LandSlope, HouseStyle(층 수), RoofStyle, RoofMatl, Exterior1st, Exterior2nd(외벽), MasVnrType, MasVnrArea, Foundation(기초 유형),
  - 지하: BsmtFinSF1, BsmtFinSF2(finished area square feet), BsmtUnfSF(미완성 지하실 평방 피트), TotalBsmtSF, BsmtFullBath, BsmtHalfBath,
  - 면적: LotArea, 1stFlrSF, 2ndFlrSF, GrLivArea(지상 총 면적), GarageArea, WoodDeckSF, OpenPorchSF, EnclosedPorch, 3SsnPorch, ScreenPorch, PoolArea,
  - 지상 시설: FullBath, HalfBath, Bedroom, Kitchen, TotRmsAbvGrd(욕실 제외하고 지상의 총 방 개수), Fireplaces, GarageType, GarageYrBlt, GarageFinish, GarageCars,
  - 이외 기타: MiscFeature, MiscVal(misc features들의 가치($)),

- 속성 정보: Utilities, BldgType(거주 타입), YearBuilt, YearRemodAdd
  - 에너지: Heating, CentralAir, Electrical

- 평가정보:
  - 건축 퀄리티 평가: BsmtFinType1(basement finished areaBsmtFinType2), OverallQual(마감, 재료 평가점수), OverallCond, ExterQual, ExterCond(현재 외벽 자재 상태), BsmtQual(지하높이 평가점수), BsmtCond
  - 속성 관련 평가: HeatingQC,
  - 시설 관련 평가: BsmtExposure, KitchenQual, FireplaceQu, GarageQual, GarageCond, PoolQC, Fence(Quality)

- 판매 정보: MoSold, YrSold, SaleType, SaleCondition, SalePrice

# EDA (1) 데이터 구조 파악

## 범주형 변수 vs 연속형 변수의 구분

범주형 변수:

MSSubClass
MSZoning
Street
Alley
LotShape
LandContour
Utilities
LotConfig
LandSlope
Neighborhood
Condition1, 2
BldgType
HouseStyle
OverallQual
OverallCond
RoofStyle
RoofMatl
Exterior1st, 2nd
MasVnrType
ExterQual
ExterCond
Foundation
BsmtQual
BsmtCond
BsmtExposure
BsmtFinType1, 2
Heating
HeatingQC
CentralAir
KitchenQual
Functional
FireplaceQu
GarageType
GarageFinish
GarageQual
GarageCond
PavedDrive
PoolQC
Fence
MiscFeature
SaleType
SaleCondition

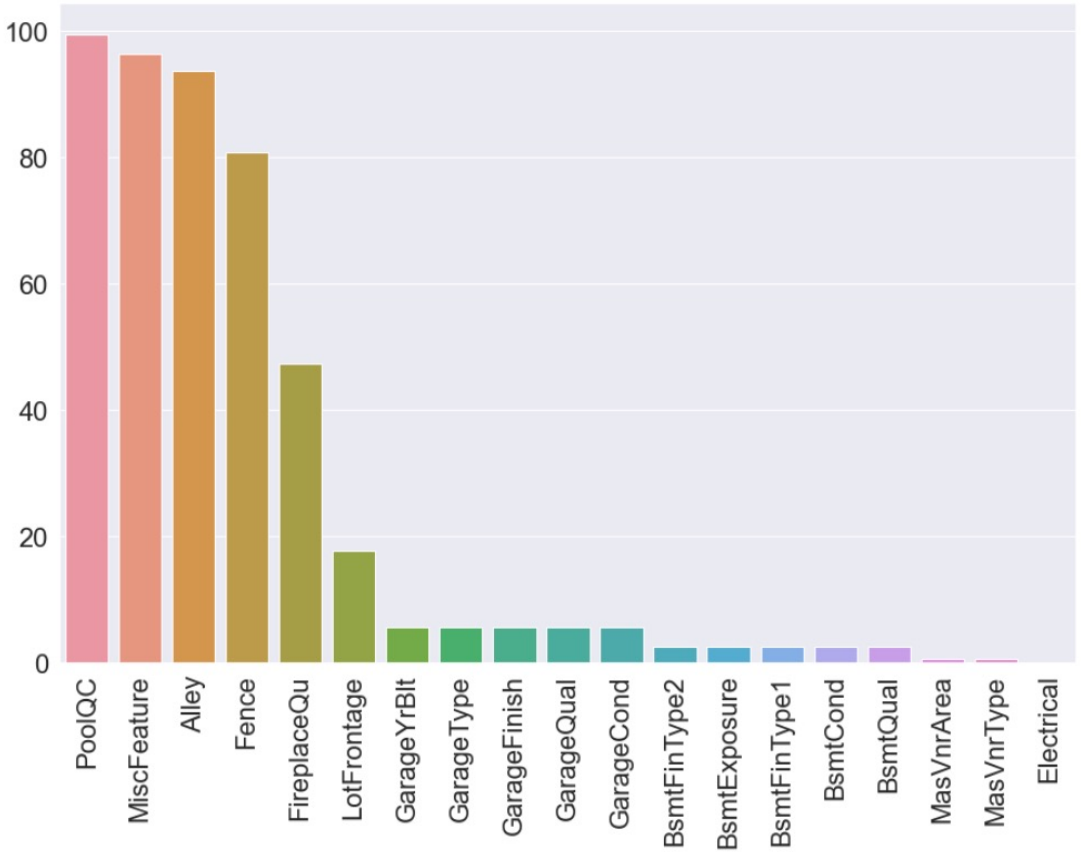
연속형 변수:

LotFrontage
LotArea
YearBuilt
YearRemodAdd
MasVnrArea
BsmtFinSF1, 2
BsmtUnfSF
TotalBsmtSF
1st, 2nd FlrSF
LowQualFinSF
GrLivArea
BsmtFullBath, HalfBath
FullBath, HalfBath
Bedroom
Kitchen
TotRmsAbvGrd
Fireplaces
GarageYrBlt
GarageCars
GarageArea
WoodDeckSF
OpenPorchSF
EncloosedPorch
3SsnPorch
ScreenProch
PoolArea
MiscVal
MoSold
YrSold

범주형 변수이지만 숫자로 나타난 변수들: MSSubClass, OverallQual, OverallCond, YrSold, MoSold는 string 형태로 변환

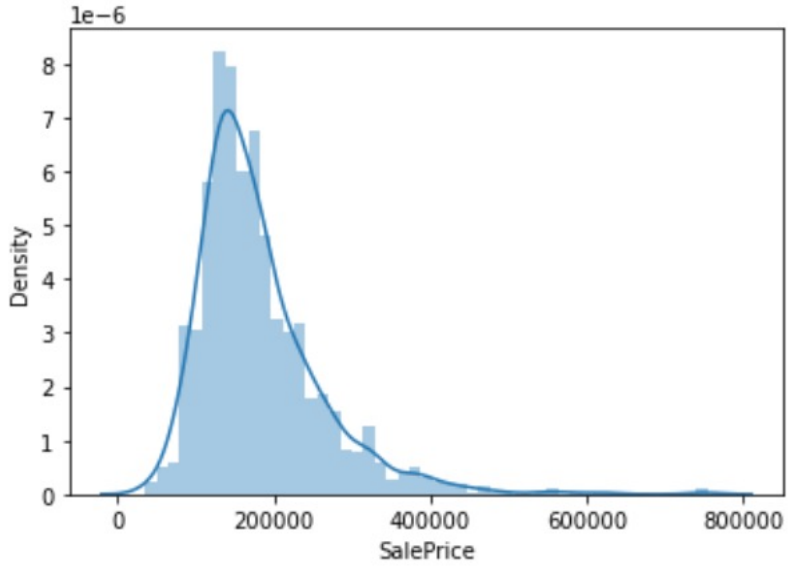
# EDA (1) 데이터 구조 파악

▶결측률 확인(Missing Ratio)




나타난 결측치 중, 범주형 변수는 'None'으로, 연속형 변수 중 'LotFrontage'는 'Neighbor' 열을 이용해 지역별 중앙값으로, 나머지 연속형 변수들은 0으로 채워줌

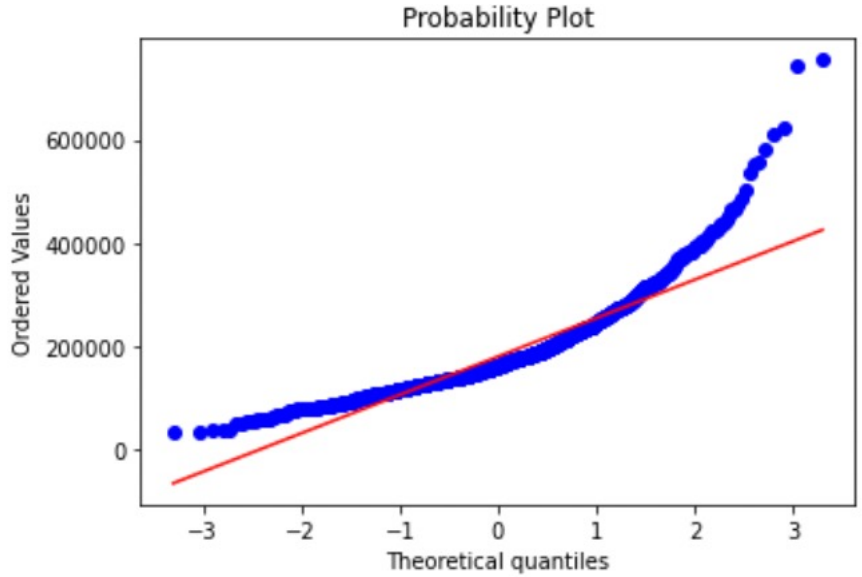
# EDA (1) 데이터 구조 파악



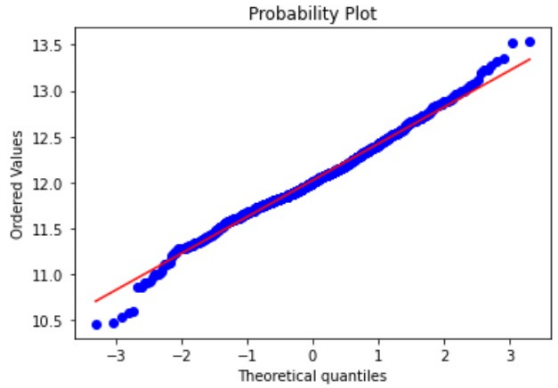

```
In [70]: data['SalePrice'].describe()

Out[70]: count      1460.000000
         mean     180921.195890
         std       79442.502883
         min       34900.000000
         25%      129975.000000
         50%      163000.000000
         75%      214000.000000
         max      755000.000000
         Name: SalePrice, dtype: float64
```

► SalePrice 의 분포 확인

►Q-Q plot 이용한 SalePrice 정규성 검정



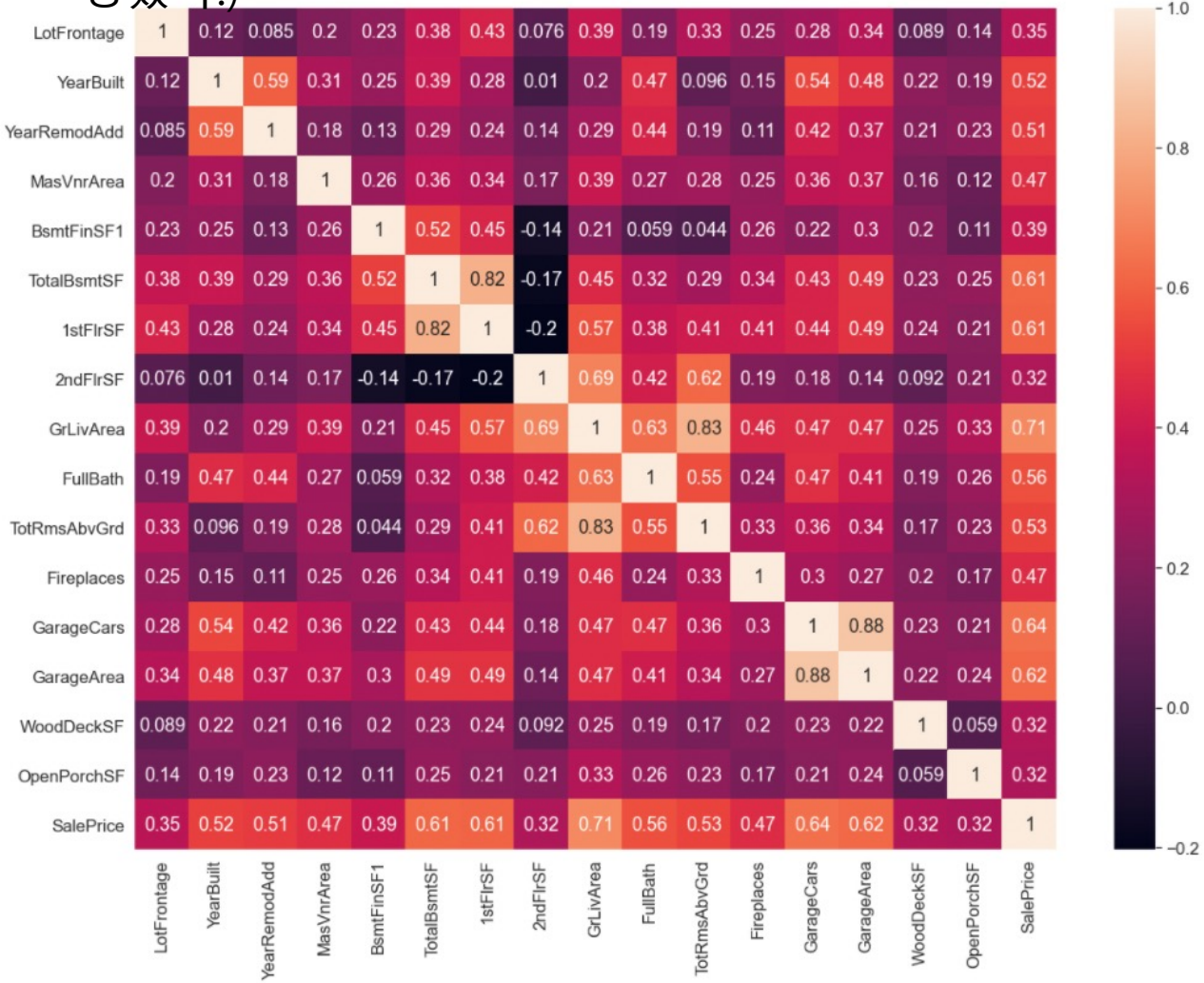

```
In [43]: #로그 변환(멱변환)
         data['SalePrice'] = np.log1p(data['SalePrice'])
         stats.probplot(data['SalePrice'], plot=plt)

Out[43]: ((array([-3.30513952, -3.04793228, -2.90489705, ...,  2.90489705,
                   3.04793228,  3.30513952]),
           array([10.46027076, 10.47197813, 10.54273278, ..., 13.34550853,
                  13.52114084, 13.53447435])),
          (0.398259646654151, 12.024057394918403, 0.9953761551826701))
```




'SalePrice'의 분포가 정규성을 띄지 않아, 정규성을 띄도록 로그 변환

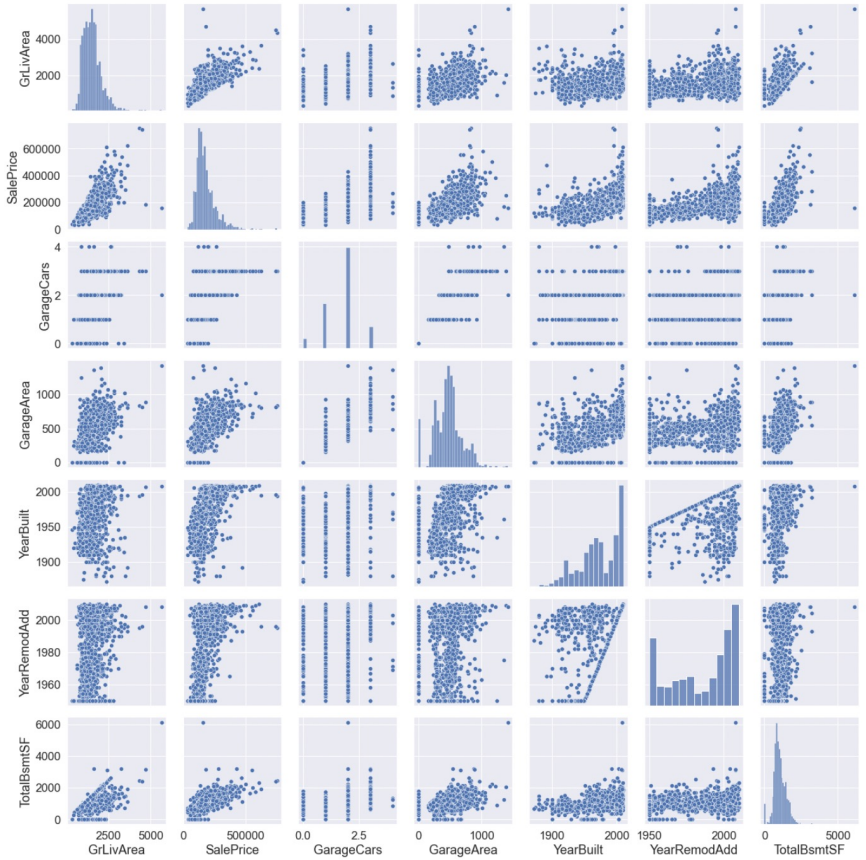
# EDA (2) 변수 관계 파악

►HeatMap을 이용한 상관 관계 파악(단, 변수간 상관 계수가 0.3이상인 변수들로만 구성했다.)



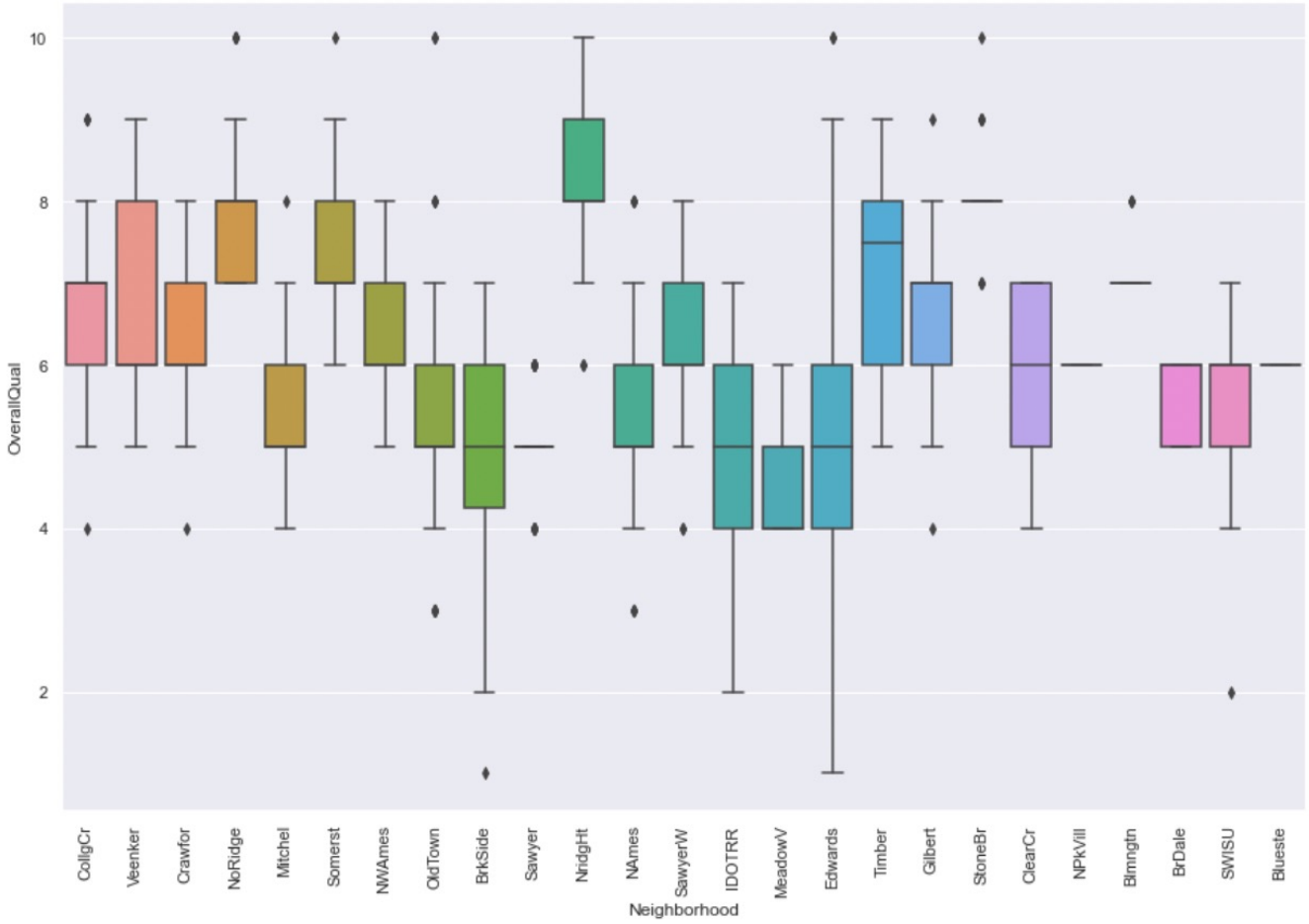

<유의미한 상관관계>
(0.71) GrLivArea – SalePrice: 주택 면적과 가격은 유의미한 양의 상관관계가 있는 것으로 파악
(0.64) GarageCars – SalePrice
(0.62) GarageArea – SalePrice
(0.61) TotalBsmtSF – SalePrice
(0.59) YearBuilt – YearRemodAdd: 두 수가 같으면 remodel이 이루어 지지 않은 것.
(0.54) YearBuilt – GarageCars
(0.52) YearBuilt – SalePrice

대부분 'SalePrice' 변수와 유의미한 상관관계가 나타난다.
SalePrice를 중심으로 한 분석을 진행해볼 수 있을 것으로 예상



# EDA (2) 변수 관계 파악

▶지역 별 OverallQuality Box-Plot으로 확인




지역별로 overall quality 점수의 편차를 확인해볼 수 있다.
가장 quality 점수가 높게 나타나는 곳은 'NridgHt'이고 'Edward'가 최대 최소 간의 격차가 가장 크게 나타난다.

지역별로 주택의 특성을 비교해보면서 'overallquality'에 영향을 미치는 요인을 파악할 수 있을 것으로 보인다.

# 분석 과제 리스트업

1. 집 값 예측: 집 값에 큰 영향 미치는 변수들로 데이터 재구성하여 집 값 예측하는 모델 만들기

2. 평가 정보 예측(Quality 관련 변수들): quality 변수 중 하나를 정해서, 다른 변수를 바탕으로 예측할 수 있도록 함