3주차 과제

김유진

21.08.15

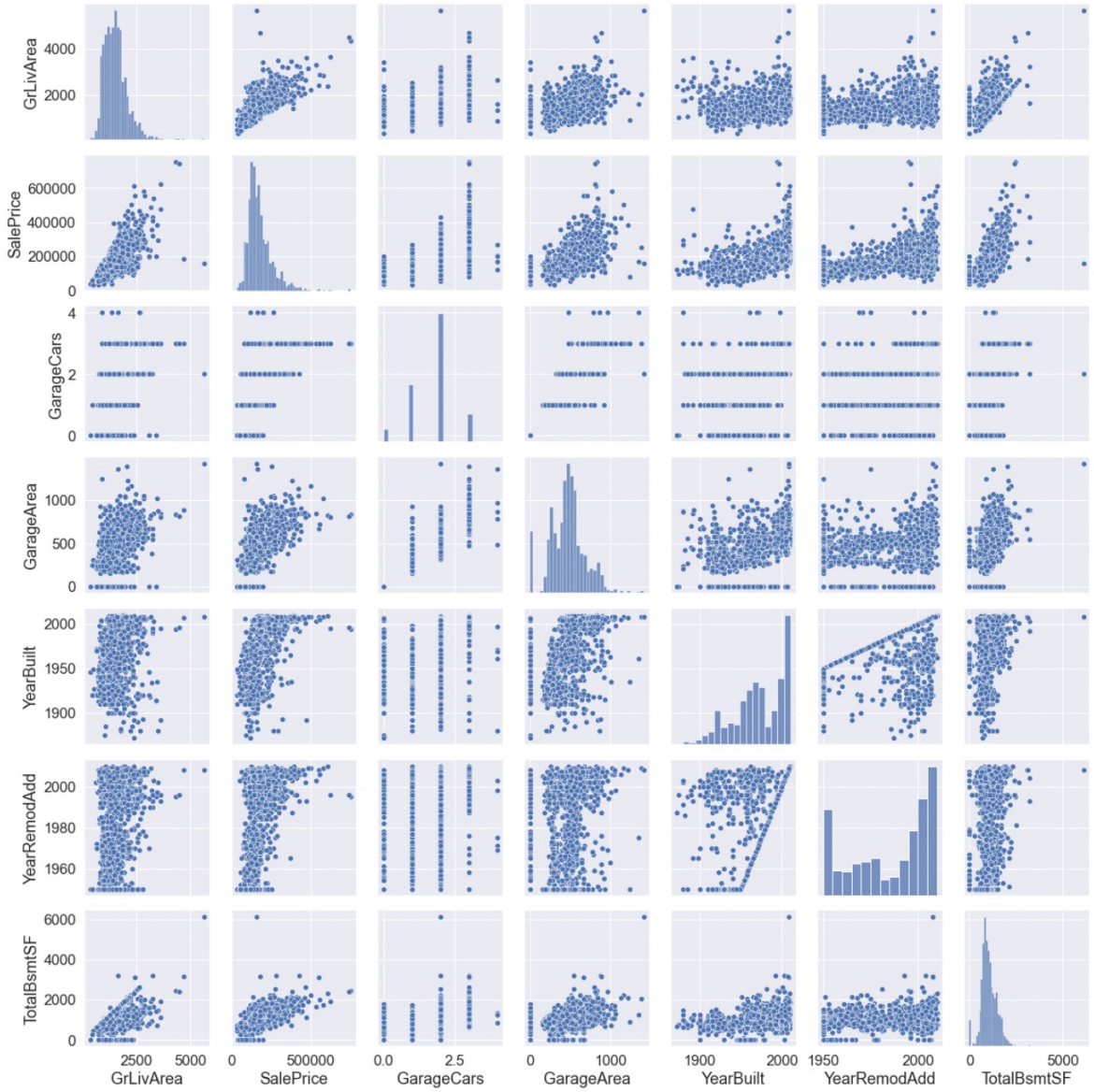
설정 주제: 집 값 예측 모델 설계 (python 이용)
- Feature extraction

1. Continuous Features

<1주차 내용 - 유의미한 상관관계>

SalePrice와 유의미한 상관관계를 보인 변수들

GrLivArea
GarageArea
YearBuilt
TotalBsmtSF
GarageCars

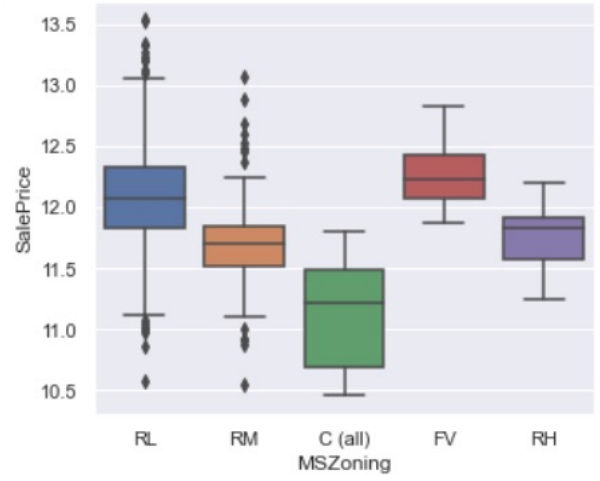


설정 주제: 집 값 예측 모델 설계 (python 이용)
- Feature extraction

2. Categorial Features

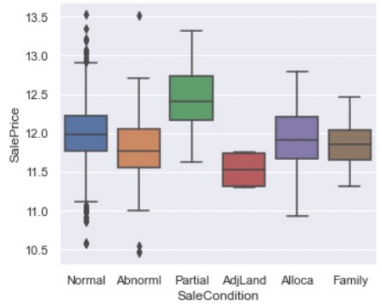
MSZoning, Street, Alley, LotShape, LandContour, Utilities, LotConfig, LandSlope, Neighborhood, Condition1, Condition2, BldgType, HouseStyle, RoofStyle, RoofMatl, Exterior1st, Exterior2nd, MasVnrType, ExterQual, ExterCond, Foundation, BsmtQual, BsmtCond, BsmtExposure, BsmtFinType1, BsmtFinType2, Heating, HeatingQC, CentralAir, KitchenQual, Functional, FireplaceQu, GarageType, GarageFinish, GarageQual, GarageCond, PavedDrive, PoolQC, Fence, MiscFeature, SaleType, SaleCondition

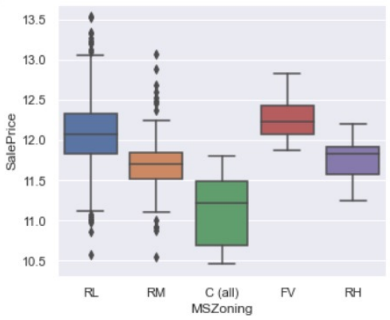
SalePrice와 boxplot으로 상관관계 살펴보기

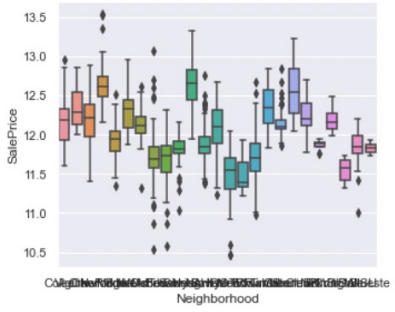


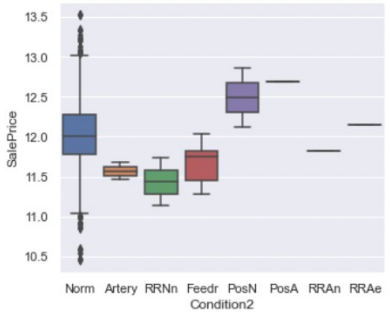
설정 주제: 집 값 예측 모델 설계 (python 이용)
- Feature extraction

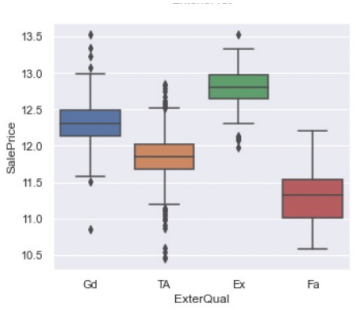
2. Categorial Features – SalePrice에 영향을 많이 끼친다고 판단되는 변수들
'MSZoning', 'Neighborhood', 'Condition2', 'ExterQual', 'BsmtQual', 'Heating', 'CentralAir', 'KitchenQual',
'GarageType', 'GarageQual', 'PoolQC', 'SaleType', 'SaleCondition'

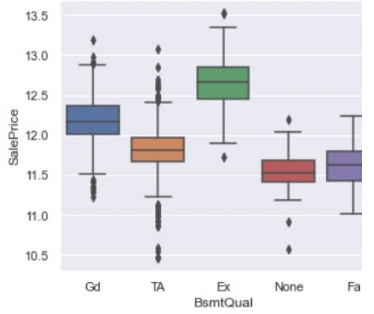


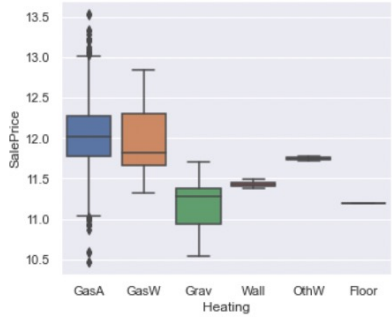


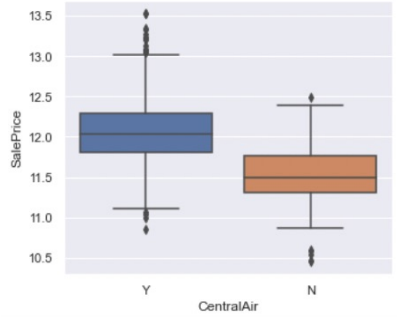


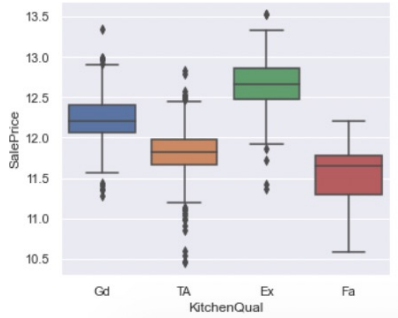


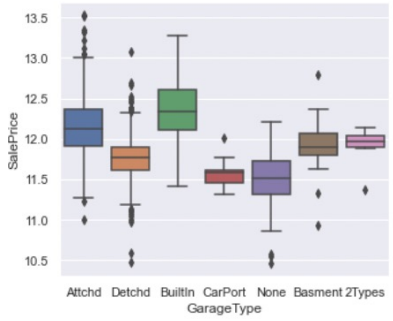


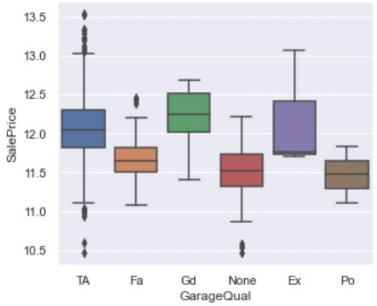


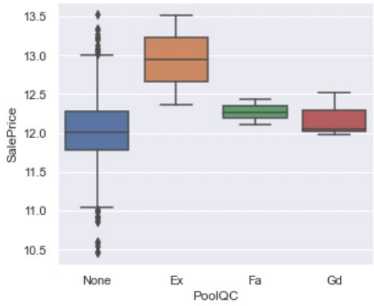


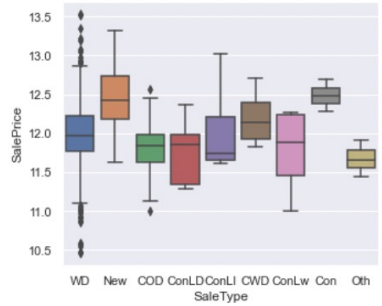












설정 주제: 집 값 예측 모델 설계 (python 이용)
- Feature extraction

Continuous + Categorial 변수들을 추려 18개의 column으로 구성된 새로운 데이터 프레임 구성

	GrLivArea	GarageArea	YearBuilt	TotalBsmtSF	GarageCars	MSZoning	Neighborhood	Condition2	ExterQual	BsmtQual	Heating	CentralAir	KitchenQua
0	1710	548	2003	856	2	RL	CollgCr	Norm	Gd	Gd	GasA	Y	Gd
1	1262	460	1976	1262	2	RL	Veenker	Norm	TA	Gd	GasA	Y	TA
2	1786	608	2001	920	2	RL	CollgCr	Norm	Gd	Gd	GasA	Y	Gd
3	1717	642	1915	756	3	RL	Crawfor	Norm	TA	TA	GasA	Y	Gd
4	2198	836	2000	1145	3	RL	NoRidge	Norm	Gd	Gd	GasA	Y	Gd
...	...	...	...	...	...	...	...	...	...	...	...	...	...
1455	1647	460	1999	953	2	RL	Gilbert	Norm	TA	Gd	GasA	Y	TA
1456	2073	500	1978	1542	2	RL	NWAmes	Norm	TA	Gd	GasA	Y	TA
1457	2340	252	1941	1152	1	RL	Crawfor	Norm	Ex	TA	GasA	Y	Gd
1458	1078	240	1950	1078	1	RL	NAmes	Norm	TA	TA	GasA	Y	Gd
1459	1256	276	1965	1256	1	RL	Edwards	Norm	Gd	TA	GasA	Y	TA

1460 rows × 18 columns

설정 주제: 집 값 예측 모델 설계 (python 이용)
- Integer Encoding

Categorial 변수들을 모델에 넣기 위해 정수로 인코딩
(SalePrice를 기준으로 평균값을 확인하고 분류)

```
In [56]: #categorial 변수들을 숫자로 바꾸기 위해 평균값 살펴보기

for cat in strong_categorial:
    g = new_df.groupby(cat)["SalePrice"].mean()
    print(g)
```

```
MSZoning
C (all)    11.118275
FV         12.246621
RH         11.749848
RL         12.085891
RM         11.692901
Name: SalePrice, dtype: float64
Neighborhood
Blmngtn    12.169421
Blueste    11.826543
BrDale     11.547874
BrkSide    11.679736
ClearCr    12.239905
CollgCr    12.163647
Crawfor    12.206664
Edwards    11.712321
Gilbert    12.155809
IDOTRR     11.446901
MeadowV    11.474533
Mitchel    11.933954
NAmes      11.868052
```

MSZ_num	NbHd_num	Cond2_num	ExtQ_num	BsQ_num	CA_num	KiQ_num	GT_num	GQ_num	Pool_num	SlTy_num	S
3	3	2	3	0	1	3	3	4	1	2	
3	3	2	2	0	1	2	3	4	1	2	
3	3	2	3	0	1	3	3	4	1	2	
3	3	2	2	0	1	3	1	4	1	2	
3	3	2	3	0	1	3	3	4	1	2	
...	...	...	...	...	...	...	...	...	...	...	
3	3	2	2	0	1	2	3	4	1	2	
3	3	2	2	0	1	2	3	4	1	2	
3	3	2	4	0	1	3	3	4	1	2	
3	2	2	2	0	1	3	3	4	1	2	
3	2	2	3	0	1	2	3	4	1	2	

_num 이란 이름으로 컬럼을 만들어 인코딩한 변수들을 저장

설정 주제: 집 값 예측 모델 설계 (python 이용)
- XGBoost 사용하기

판단 이유)
1) 1460row로 데이터가 많지 않아 과적합이 생길 수 있는데, 이를 방지해줄 수 있는 장치가 있다.
2) CART 앙상블 모델 사용 -> 회귀 시도 가능

따라서, XGBoost로 시도해보기로 결정

그러나, 사용 중인 m1 맥에서 xgboost 사용이 안됨

```
XGBoostError: XGBoost Library (libxgboost.dylib) could not be loaded.
Likely causes:
  * OpenMP runtime is not installed (vcomp140.dll or libgomp-1.dll for Windows, libomp.dylib for Mac OSX, libgomp.so
for Linux and other UNIX-like OSes). Mac OSX users: Run `brew install libomp` to install OpenMP runtime.
  * You are running 32-bit Python on a 64-bit OS
Error message(s): ['dlopen(/Users/yujinkim/opt/anaconda3/lib/python3.8/site-packages/xgboost/lib/libxgboost.dylib,
6): Library not loaded: /usr/local/opt/libomp/lib/libomp.dylib\n  Referenced from: /Users/yujinkim/opt/anaconda3/lib/
python3.8/site-packages/xgboost/lib/libxgboost.dylib\n  Reason: image not found']
```

설정 주제: 집 값 예측 모델 설계 (python 이용)

- 다른 회귀 모델

LinearRegression, Lasso, Ridge 시도

```
#Linear Regression, Lasso, Ridge 시도
linear = LinearRegression()
lasso = Lasso()
ridge = Ridge()

#학습
linear.fit(X_train, Y_train)
lasso.fit(X_train,Y_train)
ridge.fit(X_train, Y_train)
```

모델별 RMSE 계산

```
In [18]: models = [linear, lasso, ridge]
         get_rmses(models)

         LinearRegression RMSE: 0.157
         Lasso RMSE: 0.189
         Ridge RMSE: 0.158
```