코멘토 직무부트캠프_1주차

양재영

2021.04.11

No. 1 데이터 소개(명목형)

No	Variable Name	Type	Description	Missing Value	Mode Value
1	MSSubClass	명목형	판매와 관련된 주거 시설 유형	0	*20*
2	MSZoning	명목형	판매의 일반적인 구역 분류	0	*RL*
3	Street	명목형	건물과 연결된 도로의 유형	0	*Pave*
4	Alley	명목형	건물과 연결된 골목의 유형	1369	*NA*
5	LotShape	명목형	건물의 일반적인 모양	0	*Reg*
6	LandContour	명목형	건물의 평탄함 정도	0	*Lvl*
7	Utilities	명목형	사용가능한 공익사업의 유형	0	*AllPub*
8	LotConfig	명목형	Lot 배치(구성)	0	*Inside*
9	LandSlope	명목형	건물의 경사	0	*Gtl*
10	Neighborhood	명목형	Ames 도시 내 물리적 위치	0	*NAmes*
11	Condition1	명목형	다양한 조건에 대한 근접성	0	*Norm*
12	Condition2	명목형	다양한 조건에 대한 근접성(둘 이상의 경우)	0	*Norm*
13	BldgType	명목형	주거 형태	0	*1Fam*
14	HouseStyle	명목형	주거 스타일	0	*1Story*
15	RoofStyle	명목형	지붕의 유형	0	*Gable*
16	RoofMatl	명목형	지붕의 재료	0	*CompShg*
17	Exterior1st	명목형	건물의 외부 덮개	0	*VinylSd*

No	Variable Name	Type	Description	Missing Value	Mode Value
18	Exterior2nd	명목형	건물의 외부 덮개	0	*VinylSd*
19	MasVnrType	명목형	Masonry veneer 유형	8	*None*
20	Foundation	명목형	건물의 토대 유형	0	*PConc*
21	BsmtExposure	명목형	벽이나 정원의 도보 적합도	38	*No*
22	BsmtFinType1	명목형	지하실 완성 면적 등급	37	*Unf*
23	BsmtFinType2	명목형	지하실 완성 면적 등급(다수일 경우)	38	*Unf*
24	Heating	명목형	난방 유형	0	*GasA*
25	CentralAir	명목형	중앙 에어컨	0	*Y*
26	Electrical	명목형	전기 시스템	1	*SBrkr*
27	Functional	명목형	집의 기능	0	*Typ*
28	GarageType	명목형	차고 위치	81	*Attchd*
29	GarageFinish	명목형	차고 내부 마감	81	*Unf*
30	PavedDrive	명목형	포장 된 진입로	0	*Y*
31	Fence	명목형	울타리 품질	1179	*NA*
32	MiscFeature	명목형	다른 변수들에서 다루지 않는 기타 기능	1406	*NA*
33	SaleType	명목형	판매 유형	0	*WD*
34	SaleCondition	명목형	판매 조건	0	*Normal*

No. 2 데이터 소개(명목형)

[명목형 변수의 Barplot]

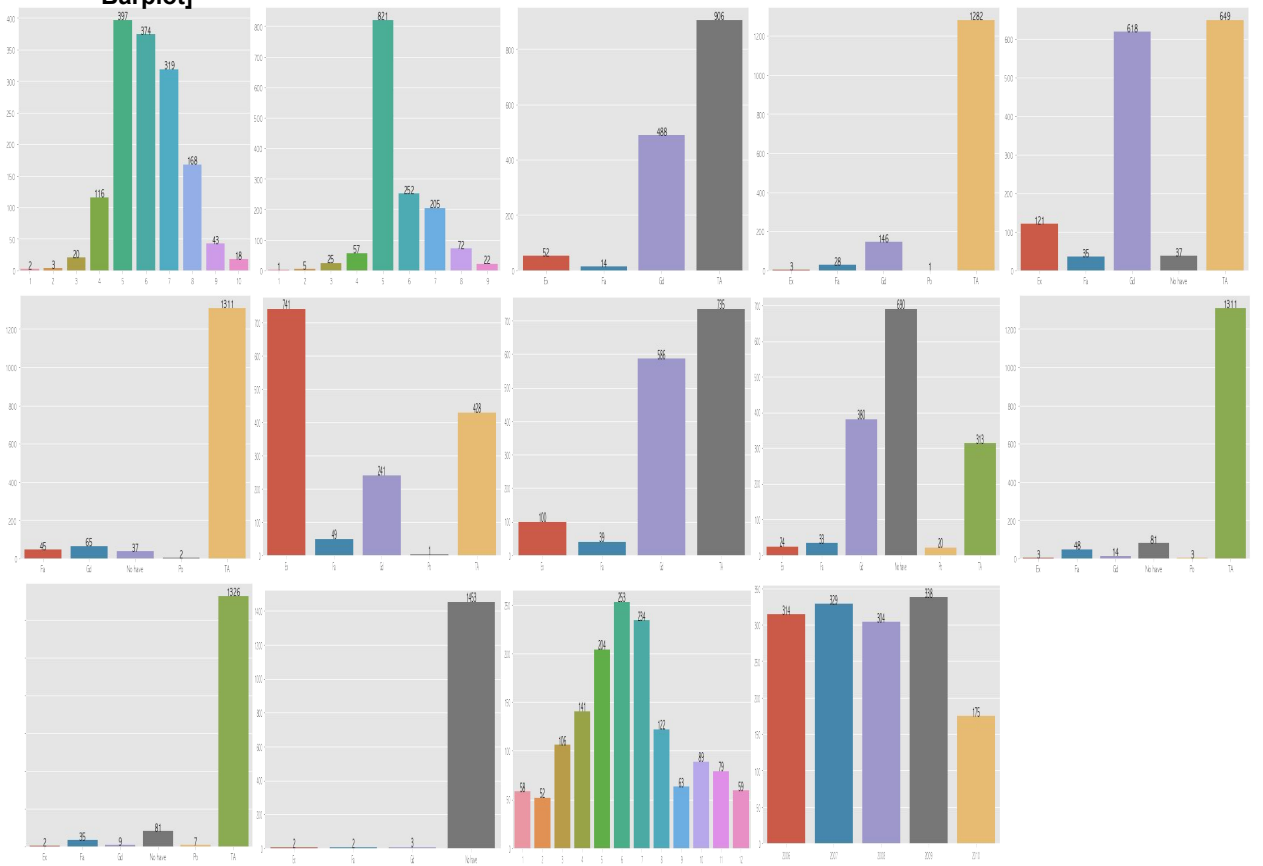
✔ Alley, MasVnrType, BsmtExposure, BsmtFinType1, BsmtFinType2, Electrical, GarageType, GarageFinish, Fence, MiscFeature 총 107개의 명목형 변수에 결측치가 존재한다.

✔ 명목형 변수의 경우, 특정 범주에 집중적으로 몰려있음을 볼 수 있다. *Ex) Street, Utilities, Condition2, RoofMatl, Heating, MiscFeature*

No. 3 데이터 소개(순서형)

No	Variable Name	Type	Description	Missing Value	Mode Value
1	OverallQual	순서형	집의 전반적인 소재 및 마감 품질 평가	0	5
2	OverallCond	순서형	집의 전반적인 상태 평가	0	5
3	ExterQual	순서형	외부 소재 품질에 대한 평가	0	TA
4	ExterCond	순서형	외부 소재의 현재 상태에 대한 평가	0	TA
5	BsmtQual	순서형	지하 높이 평가	37	TA
6	BsmtCond	순서형	지하 전반적인 상태 평가	37	TA
7	HeatingQC	순서형	난방의 품질과 상태	0	Ex
8	KitchenQual	순서형	주방 품질	0	TA
9	FireplaceQu	순서형	벽난로 품질	690	NA
10	GarageQual	순서형	차고 품질	81	TA
11	GarageCond	순서형	차고 상태	81	TA
12	PoolQC	순서형	수영장 품질	1453	NA
13	YearBuilt	순서형	건축년도	0	2006
14	YearRemodAdd	순서형	Remodeling 연도	0	1950
15	GarageYrBlt	순서형	차고가 건설된 년도	81	NA
16	MoSold	순서형	판매 된 달	0	6
17	YrSold	순서형	판매 된 연도	0	2009

[순서형 변수의 Barplot]



✔BsmtQual, BsmtCond, FireplaceQU, GarageQual, GarageCond, PoolQC, GarageYrBlt 총 7개의 순서형 변수에 결측치가 존재한다.

✔ 순서형 변수의 경우 평가와 관련된 순위척도 변수들이 대부분이며, 시각화와 최빈값을 통하여 '보통'을 의미하는 중간응답의 경향이 많음을 확인할 수 있다.

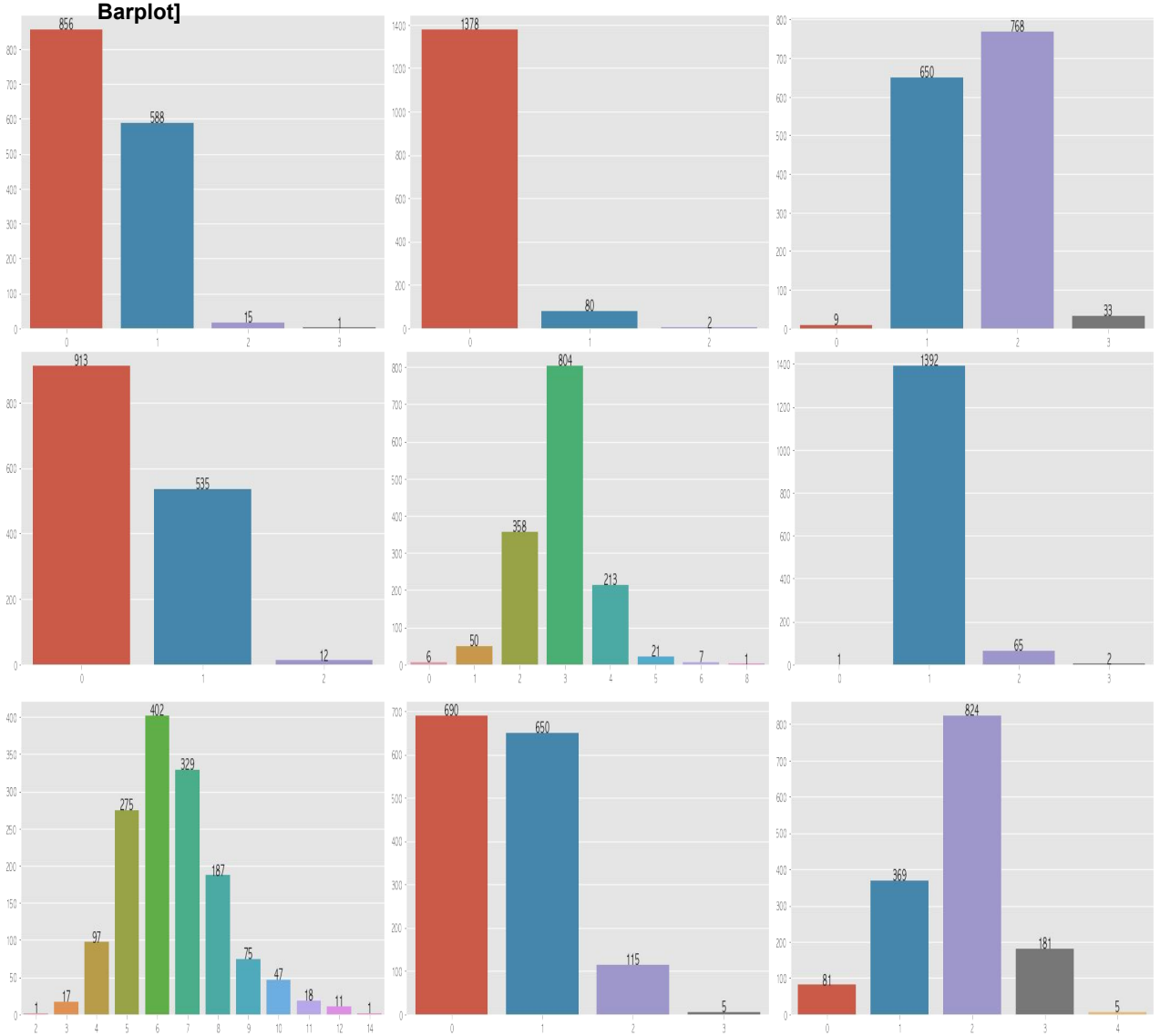
No. 4 데이터 소개(이산형)

No	Variable Name	Type	Description	Missing Value	Mode Value
1	BsmtFullBath	이산형	지하 full bathroom의 개수	0	0
2	BsmtHalfBath	이산형	지하 half bathroom의 개수	0	0
3	FullBath	이산형	지상 full bathroom의 개수	0	2
4	HalfBath	이산형	지상 half bathroom의 개수	0	0
5	BedroomAbvGr	이산형	지상 bedroom의 개수(지하 포함 X)	0	3
6	KitchenAbvGr	이산형	주방의 수	0	1
7	TotRmsAbvGrd	이산형	전체 방의 수(bathroom 포함 X)	0	6
8	Fireplaces	이산형	벽난로 개수	0	0
9	GarageCars	이산형	차고에 수용가능한 차의 수	0	2

✔ 이산형 변수에서는 결측치가 존재하지 않는다.

✔ 이산형 변수의 모든 변수는 해당 시설에 대한 개수를 나타냄을 확인할 수 있다.

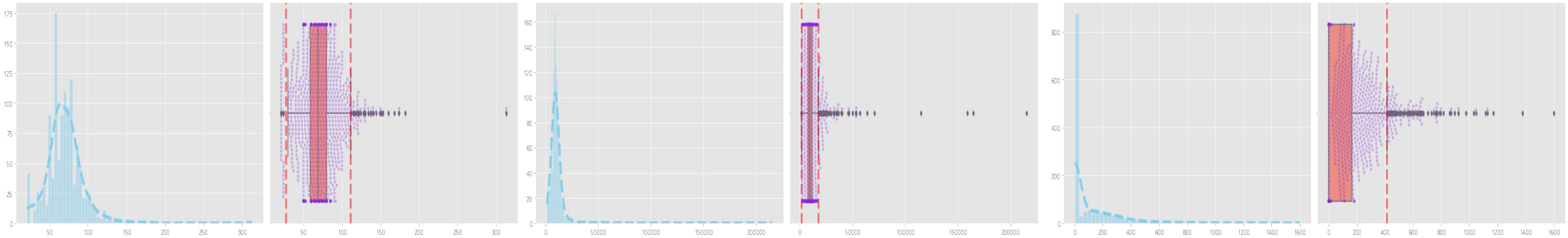
[이산형 변수의 Barplot]



No. 5 데이터 소개(연속형)

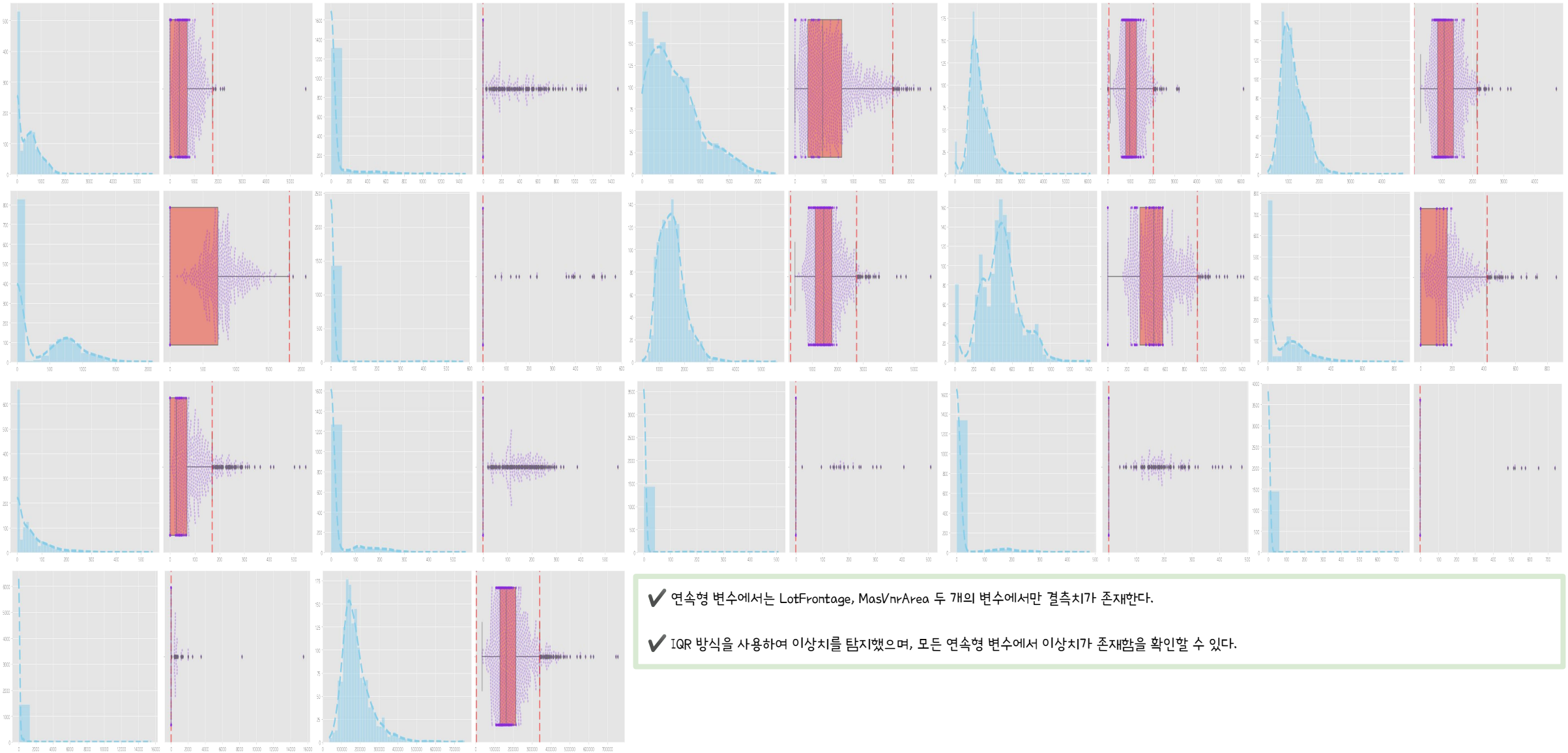
No	Variable Name	Type	Description	Missing Value	Outliers	Mean	Median
1	LotFrontage	연속형	건물과 연결된 도로의 직선 거리	259	88	70.0	69.0
2	LotArea	연속형	집의 크기	0	69	10516.8	9478.5
3	MasVnrArea	연속형	Masonry veneer 구역의 면적	8	96	103.7	0.0
4	BsmtFinSF1	연속형	완성된 지하의 크기	0	7	443.6	383.5
5	BsmtFinSF2	연속형	완성된 지하의 크기2	0	167	46.5	0.0
6	BsmtUnfSF	연속형	완성되지 않은 지하의 크기	0	29	567.2	477.5
7	TotalBsmtSF	연속형	지하 전체 크기	0	61	1057.4	991.5
8	1stFlrSF	연속형	1층의 크기	0	20	1162.6	1087.0
9	2ndFlrSF	연속형	2층의 크기	0	2	347.0	0.0
10	LowQualFinSF	연속형	저품질 마감 크기	0	26	5.8	0.0

No	Variable Name	Type	Description	Missing Value	Outliers	Mean	Median
11	GrLivArea	연속형	생활 공간의 면적	0	31	1515.5	1464.0
12	GarageArea	연속형	차고의 면적	0	21	473.0	480.0
13	WoodDeckSF	연속형	Wood deck의 면적	0	32	94.2	0.0
14	OpenPorchSF	연속형	Open porch의 면적	0	77	46.7	25.0
15	EnclosedPorch	연속형	Enclosed porch의 면적	0	208	22.0	0.0
16	3SsnPorch	연속형	Three season porch의 면적	0	24	3.4	0.0
17	ScreenPorch	연속형	Screen porch의 면적	0	116	15.1	0.0
18	PoolArea	연속형	수영장 면적	0	7	2.8	0.0
19	MiscVal	연속형	기타 기능의 가치	0	52	43.5	0.0
20	SalePrice	연속형	판매 가격	0	61	180921.2	163000.0



No. 6 데이터 소개(연속형)

[연속형 변수의 Histogram & Boxplot]

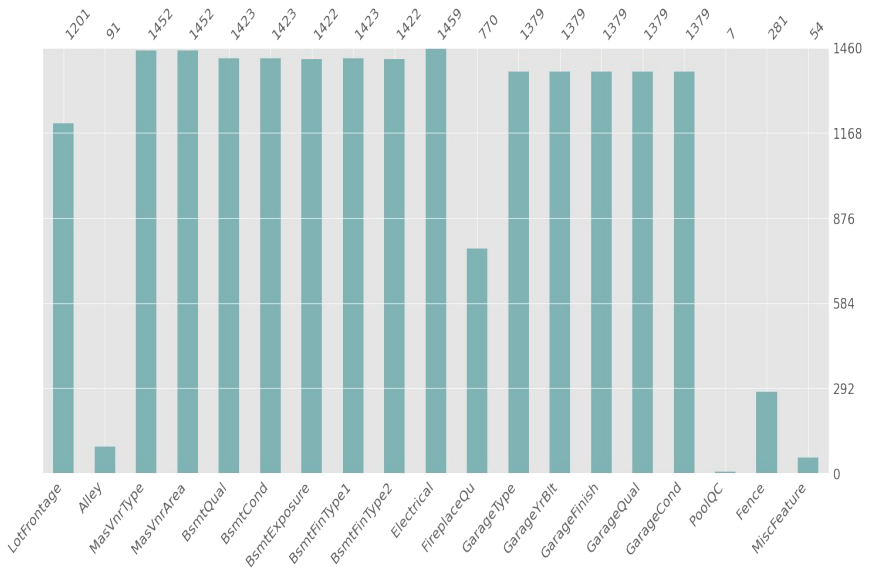


✔ 연속형 변수에서는 LotFrontage, MasVnrArea 두 개의 변수에서만 결측치가 존재한다.

✔ IQR 방식을 사용하여 이상치를 탐지했으며, 모든 연속형 변수에서 이상치가 존재함을 확인할 수 있다.

No. 7 데이터 결측치 처리

[원본 데이터의 결측치 개수 시각화]



[변수에 해당하는 시설이 없는 경우 결측치 수정 코드]

```
object_list = ['Alley', 'BsmtQual', 'BsmtCond', 'BsmtExposure', 'BsmtFinType1', 'BsmtFinType2',
               'FireplaceQu', 'GarageType', 'GarageFinish', 'GarageQual', 'GarageCond',
               'PoolQC', 'Fence', 'MiscFeature']

# 원래 데이터를 보존하기 위해 새로운 DataFrame 생성
modify_data = raw_data.copy()

# 결측치를 의미하지 않는 값에 'No have' 채워넣기
for feature in object_list:
    null_data = modify_data[feature].isnull().sum()
    print('Number of null values of {} *Before Modify* :'.format(feature), null_data, '\n')

    modify_data.loc[modify_data[feature].isnull(), feature] = 'No have'

    null_data = modify_data[feature].isnull().sum()
    print('Number of null values of {} *After Modify* :'.format(feature), null_data)
    print('-' * 60)
```

[LotFrontage 변수의 결측치 수정 코드]

```
# 'LotFrontage' 결측치를 대체해주는 함수 만들기
def imputer_lotfrontage(raw_data):

    print('Number of null values of LotFrontage *Before Modify* :', raw_data['LotFrontage'].isnull().sum())

    # 'KNNImputer' 라이브러리 불러오기
    from sklearn.impute import KNNImputer

    # 'LotFrontage' 변수에 결측치를 가지는 'Id' 추출 후 저장
    miss_id = raw_data[raw_data['LotFrontage'].isnull()]['Id'].unique()

    # 'Id', 'LotFrontage', 'LotArea' 변수 추출 후 저장
    data = raw_data[['Id', 'LotFrontage', 'LotArea']]

    # 'LotFrontage' 변수의 값이 300 넘는 관측치를 삭제하기로 했으므로 삭제
    outlier_id = data[data['LotFrontage'] > 300]['Id']
    data = data[~data['Id'].isin(outlier_id)]

    # 'Id' 변수를 인덱스로 지정
    data = data.set_index('Id')

    # 'LotArea' 변수를 로그화 수행 후, 변수 삭제
    data['LotLogArea'] = data['LotArea'].apply(lambda x: np.log(x))
    data = data.drop('LotArea', axis = 1)

    # 모델 객체 생성 후 학습시키고 DataFrame으로 저장
    # 원본 데이터에 결측치로 기입되어 있는 관측치들만 수정해주기 위해 DataFrame을 만들 때, 인덱스를 위 데이터의 인덱스로 지정
    imputer = KNNImputer(n_neighbors = 5)
    pre_data = pd.DataFrame(imputer.fit_transform(data), columns = data.columns, index = data.index)

    # 결측치가 존재하던 관측치만 저장
    pre_data = pre_data[pre_data.index.isin(miss_id)]

    # 원본 데이터 결측치를 대입해주는 과정
    for id_value, lot_value in zip(pre_data.index, pre_data['LotFrontage']):
        for index in raw_data.index:

            if raw_data.loc[index, 'Id'] == id_value:
                raw_data.loc[index, 'LotFrontage'] = lot_value

    print('Number of null values of LotFrontage *After Modify* :', raw_data['LotFrontage'].isnull().sum())

    return raw_data
```

✔ 원본데이터의 총 19개의 변수에 결측치가 존재함을 확인할 수 있다.

✔ 하지만, Alley, BsmtQual, BsmtCond 등 14개의 명목형 변수의 결측치는 변수의 설명서를 통하여 실제 결측치가 아닌 해당 시설이 없다는 것을 알 수 있다. 따라서, 왼쪽의 과정을 통하여 14개의 변수의 결측치에 '해당 시설이 없다'라는 의미로 'No have' 값을 채워주었다.

✔ LotFrontage 변수의 결측치는 약 20%를 차지하기 때문에 예측을 통해 결측치를 대체하는 방법을 선택하였다. RadomForest 모델을 구축하여 LotFrontage 변수를 예측하는데 있어서 LotArea 변수가 중요함을 확인하였고, 데이터의 개수가 적어 KNN 알고리즘을 사용하여 대체해주었다.

✔ Electrical 변수에는 1개의 결측치만 존재하여 최빈값으로 대체해주었으며, MasVnrType, MasVnrArea 변수에는 8개의 결측치가 존재하여 각각 'None'과 '0'으로 대체해주었다.

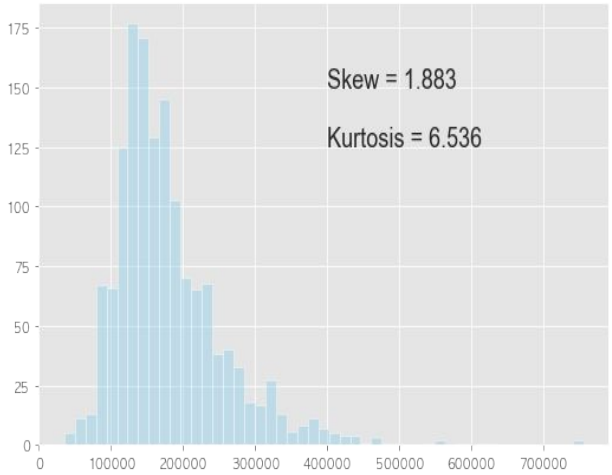
✔ 또한, GarageYrBlt 변수는 차고가 건축된 년도를 의미하며 차고가 없는 관측치에 결측치로 기재되어 있음을 확인하였고, 건축되지 않았다는 의미로 '9999'의 값으로 대체해주었다.

No. 8 목표 변수

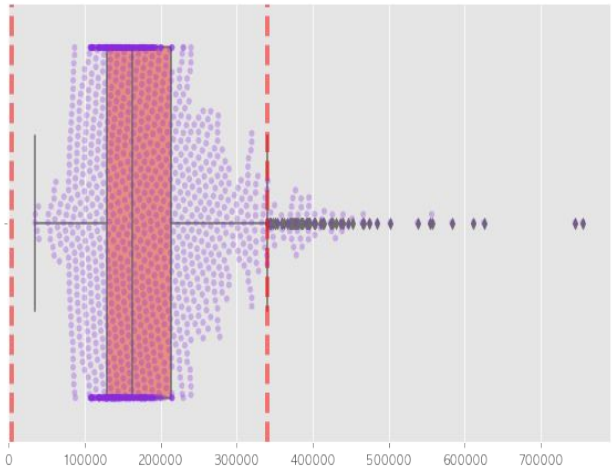
[SalePrice 요약통계량]

Mean	180921.195890
Std	79442.502883
Min	34900.000000
Max	755000.000000
25%	129975.000000
50%	163000.000000
75%	214000.000000

[Histogram of SalePrice]



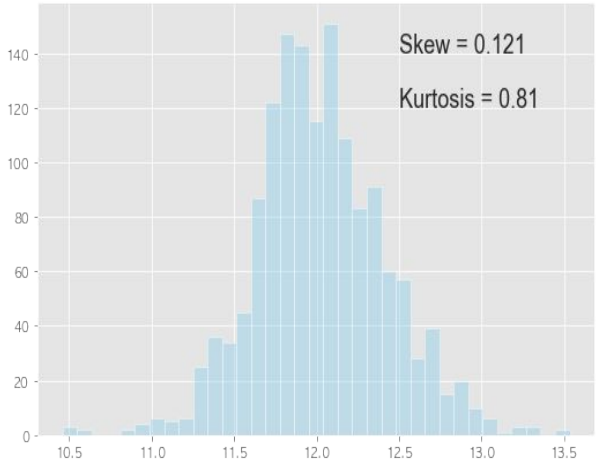
[Box & Swarm Plot of SalePrice]



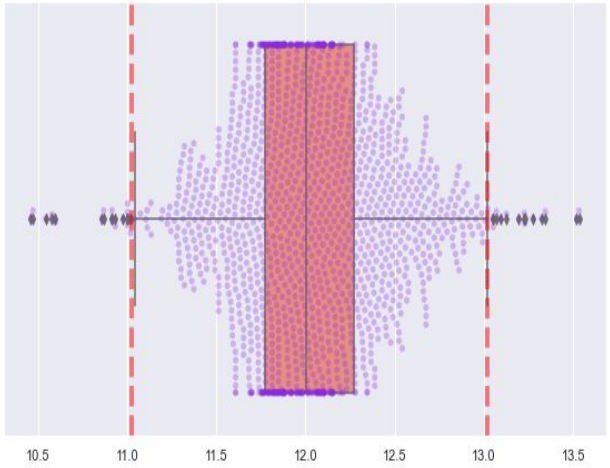
[LogSalePrice 요약통계량]

Mean	12.024051
Std	0.399452
Min	10.460242
Max	13.534473
25%	11.775097
50%	12.001505
75%	12.273731

[Histogram of LogSalePrice]



[Box & Swarm Plot of LogSalePrice]



✔'House Prices' 데이터는 2006년도 부터 2010년도 까지의 판매된 건물에 대한 정보를 담고 있으며, SalePrice 변수를 일반적으로 목표 변수로 설정할 수 있다.

✔ SalePrice 변수의 왜도는 1.833으로 왼쪽으로 치우쳐져 있는 분포를 나타내고 있으며, 100,000 ~ 200,000 사이에 많이 분포되어 있음을 확인할 수 있다.

✔ SalePrice 변수에 로그 변환을 수행하였고, 기존 데이터보다 정규분포를 따르고 있음을 확인할 수 있다.

✔ 연속형 변수와 마찬가지로 IQR 방법을 사용하여 이상치를 탐지했으며, SalePrcie와 LogSalePrice 두 변수 모두 이상치가 존재함을 확인할 수 있다.

✔ LogSalePrice 변수를 사용하는 것이 적절하다고 판단되지만, 분석을 통하여 어떤 건물이 높은 가격을 가지고 있는지 확인한 후 사용할 변수를 선택하도록 하자.

No. 9 해당 데이터를 바탕으로 진행할 수 있는 분석과제 리스트

1. 새로운 집에 대한 정보가 주어졌을 때, 해당 집이 어느 정도의 가격에 판매될지 예측하는 모델 구축

2. 가격에 이상치가 존재하기 때문에 판매 가격을 구간화 하여 분류 모델 구축

3. 주택 등급(OverallQual)에 영향을 미치는 변수를 파악 후, 주택 등급을 판단하는 분류 모델 구축

4. 차고가 건설된 년도에 따라 집값이 변화하는지 파악

5. 가설검정을 통해 가격 변수에 영향을 미치는지 파악

    1) 리모델링을 한 건물이 가격이 높을 것이다.

    2) 1층과 2층을 모두 보유하고 있는 건물이 가격이 높을 것이다.

    3) 집의 크기가 클 수록 건물의 가격이 높을 것이다.