```
# 시험환경 세팅 (코드 변경 X)
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split

def exam_data_load(df, target, id_name="", null_name=""):
    if id_name == "":
        df = df.reset_index().rename(columns={"index": "id"})
        id_name = 'id'
    else:
        id_name = id_name

    if null_name != "":
        df[df == null_name] = np.nan

    X_train, X_test = train_test_split(df, test_size=0.2, shuffle=True, random_state=2021)
    y_train = X_train[[id_name, target]]
    X_train = X_train.drop(columns=[id_name, target])
    y_test = X_test[[id_name, target]]
    X_test = X_test.drop(columns=[id_name, target])
    return X_train, X_test, y_train, y_test

df = pd.read_csv("data/house/train.csv")
X_train, X_test, y_train, y_test = exam_data_load(df, target='SalePrice', id_name='Id')

X_train.shape, X_test.shape, y_train.shape, y_test.shape
```

```
((1168, 79), (292, 79), (1168, 2), (292, 2))
```

데이터 분석

```
X_train.head()
```

	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	...
81	120	RM	32.0	4500	Pave	NaN	Reg	Lvl	AllPub	FR2	...
1418	20	RL	71.0	9204	Pave	NaN	Reg	Lvl	AllPub	Inside	...
1212	30	RL	50.0	9340	Pave	NaN	Reg	Lvl	AllPub	Inside	...
588	20	RL	65.0	25095	Pave	NaN	IR1	Low	AllPub	Inside	...
251	120	RM	44.0	4750	Pave	NaN	IR1	HLS	AllPub	Inside	...

5 rows × 79 columns

```
import pandas as pd
X_train.shape, X_test.shape
```

```
((1168, 79), (292, 79))
```

```
# X_train.info()
```

```
X_train.isnull().sum().sort_values().tail(3)
```

```
Alley          1098
MiscFeature    1124
PoolQC         1163
dtype: int64
```

심각한 결측치 제거

```
X_train.drop(columns=['PoolQC','MiscFeature','Alley'],inplace=True)
X_test.drop(columns=['PoolQC','MiscFeature','Alley'],inplace=True)
```

```
X_train.shape, X_test.shape
```

```
((1168, 76), (292, 76))
```

결측치 보간

```
num_cols=X_train.select_dtypes(['float64','int64']).columns
cat_cols=X_train.select_dtypes(['object']).columns
```

```
X_train.head()
```

	MSSubClass	MSZoning	LotFrontage	LotArea	Street	LotShape	LandContour	Utilities	LotConfig	LandSlope
81	120	RM	32.0	4500	Pave	Reg	Lvl	AllPub	FR2	Gtl
1418	20	RL	71.0	9204	Pave	Reg	Lvl	AllPub	Inside	Gtl
1212	30	RL	50.0	9340	Pave	Reg	Lvl	AllPub	Inside	Gtl
588	20	RL	65.0	25095	Pave	IR1	Low	AllPub	Inside	Sev
251	120	RM	44.0	4750	Pave	IR1	HLS	AllPub	Inside	Mod

5 rows × 76 columns

```
for col in num_cols:
    X_train[col]=X_train[col].fillna(X_train[col].mean())
    X_test[col]=X_test[col].fillna(X_test[col].mean())
```

```
X_train.isnull().sum().sum(), X_test.isnull().sum().sum()
```

```
(2584, 669)
```

```
for col in cat_cols:
    X_train[col]=X_train[col].fillna(X_train[col].mode()[0])
    X_test[col]=X_test[col].fillna(X_test[col].mode()[0])
```

```
X_train.isnull().sum().sum(), X_test.isnull().sum().sum()
```

```
(0, 0)
```

범주형 변수 원핫인코딩

```
X_train.shape, X_test.shape
```

```
((1168, 76), (292, 76))
```

```
X_train= pd.get_dummies(X_train)
X_test= pd.get_dummies(X_test)
```

```
X_train.shape, X_test.shape
```

```
((1168, 275), (292, 246))
```

train test align 작업

```
X_train, X_test = X_train.align(X_test,join='left',fill_value=0,axis=1)
```

```
X_train.shape, X_test.shape
```

```
((1168, 275), (292, 275))
```

랜덤포레스트 학습

```
from sklearn.ensemble import RandomForestRegressor
```

```
rf = RandomForestRegressor(random_state=42, n_estimators=500)
```

```
y_train.head()
```

	Id	SalePrice
81	82	153500
1418	1419	124000
1212	1213	113000
588	589	143000
251	252	235000

```
# y_train.pop('Id')
```

```
rf.fit(X_train,y_train['SalePrice'])
pred= rf.predict(X_test)
```

```
from sklearn.metrics import mean_squared_error, r2_score
```

```
rmse= mean_squared_error(y_test['SalePrice'],pred,squared=False)
r2= r2_score(y_test['SalePrice'],pred)
```

```
C:\Users\pjjeo\anaconda3\Lib\site-packages\sklearn\metrics\_regression.py:492: FutureWarning: 'squared' is deprecated in ve
  warnings.warn(
```

```
print("RMSE:", rmse)
print("R²:", r2)
```

```
RMSE: 24844.269928350943
R²: 0.8942271397027238
```