TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

KHOA KHOA HỌC MÁY TÍNH

Báo cáo kết quả
TIỀN XỬ LÝ DỮ LIỆU CHO BÀI TOÁN DỰ ĐOÁN GIÁ NHÀ

Lớp: CS116.P21

Nhóm thực hiện:

Nguyễn Thị Lý - 22520837

Hà Ngũ Long Nguyên - 22520965

Nguyễn Thu Phương - 22521167

HỒ CHÍ MINH – 04/2025

1. Xử lý dữ liệu bị khuyết:

- Tập train.csv có 1460 dòng, 81 cột.
- Thực hiện kiểm tra số lượng dữ liệu bị khuyết trên tập train:

	Missing Count	Missing %
PoolQC	1453	99.520548
MiscFeature	1406	96.301370
Alley	1369	93.767123
Fence	1179	80.753425
MasVnrType	872	59.726027
FireplaceQu	690	47.260274
LotFrontage	259	17.739726
GarageType	81	5.547945
GarageYrBlt	81	5.547945
GarageFinish	81	5.547945
GarageQual	81	5.547945
GarageCond	81	5.547945
BsmtFinType2	38	2.602740
BsmtExposure	38	2.602740
BsmtFinType1	37	2.534247
BsmtCond	37	2.534247
BsmtQual	37	2.534247
MasVnrArea	8	0.547945
Electrical	1	0.068493

Ta thấy có 4 cột *PoolQC, MiscFeature, Alley, Fence* có tỷ lệ dữ liệu thiếu lớn hơn 80%, tiến hành loại bỏ các cột này.

- Các cột còn lại bị thiếu bao gồm: 'LotFrontage', 'MasVnrType', 'MasVnrArea', 'BsmtQual', 'BsmtCond', 'BsmtExposure', 'BsmtFinType1', 'BsmtFinType2', 'Electrical', 'FireplaceQu', 'GarageType', 'GarageYrBlt', 'GarageFinish', 'GarageQual', 'GarageCond'. Trong đó, có 3 cột kiểu số: *LotFrontage, MasVnrArea và GarageYrBlt,* các cột còn lại thuộc dạng phân loại.
    + Đối với các cột kiểu số, tiến hành điền bằng **median.**

```
median_values = numeric_cols.median()
median_values
```

```
LotFrontage      69.0
MasVnrArea        0.0
GarageYrBlt    1980.0
dtype: float64
```

    + Các cột kiểu chuỗi (categorical) điền bằng giá trị **“None”**.

+ Riêng cột *Electrical* chỉ có 1 dòng bị thiếu nên điền bằng giá trị mode của cột này.

Kết quả thực hiện xử lý dữ liệu bị khuyết:

```
# Kiểm tra các giá trị thiếu trong tập dữ liệu chưa xử lý
missing_values = df.isnull().sum()

# In ra các cột có giá trị thiếu và số lượng giá trị thiếu trong mỗi cột
print(missing_values[missing_values > 0])
```

```
Series([], dtype: int64)
```

Sau khi thực hiện tiền xử lý với dữ liệu bị khuyết, tập train còn lại 77 cột và 1460 dòng:

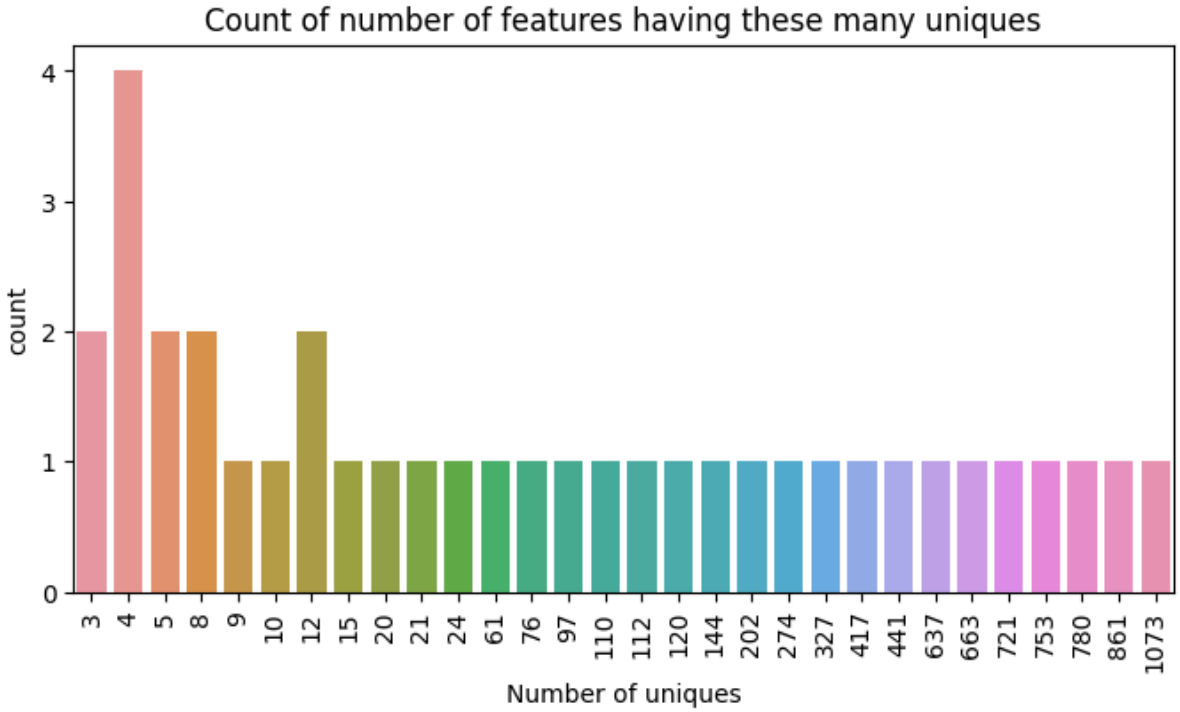
	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	LotShape	LandContour	Utilities	LotConfig	...
0	1	60	RL	65.0	8450	Pave	Reg	Lvl	AllPub	Inside	...
1	2	20	RL	80.0	9600	Pave	Reg	Lvl	AllPub	FR2	...
2	3	60	RL	68.0	11250	Pave	IR1	Lvl	AllPub	Inside	...
3	4	70	RL	60.0	9550	Pave	IR1	Lvl	AllPub	Corner	...
4	5	60	RL	84.0	14260	Pave	IR1	Lvl	AllPub	FR2	...
...	...	...	...	...	...	...	...	...	...	...	...
1455	1456	60	RL	62.0	7917	Pave	Reg	Lvl	AllPub	Inside	...
1456	1457	20	RL	85.0	13175	Pave	Reg	Lvl	AllPub	Inside	...
1457	1458	70	RL	66.0	9042	Pave	Reg	Lvl	AllPub	Inside	...
1458	1459	20	RL	68.0	9717	Pave	Reg	Lvl	AllPub	Inside	...
1459	1460	20	RL	75.0	9937	Pave	Reg	Lvl	AllPub	Inside	...

1460 rows × 77 columns

2. Xử lý outlier:

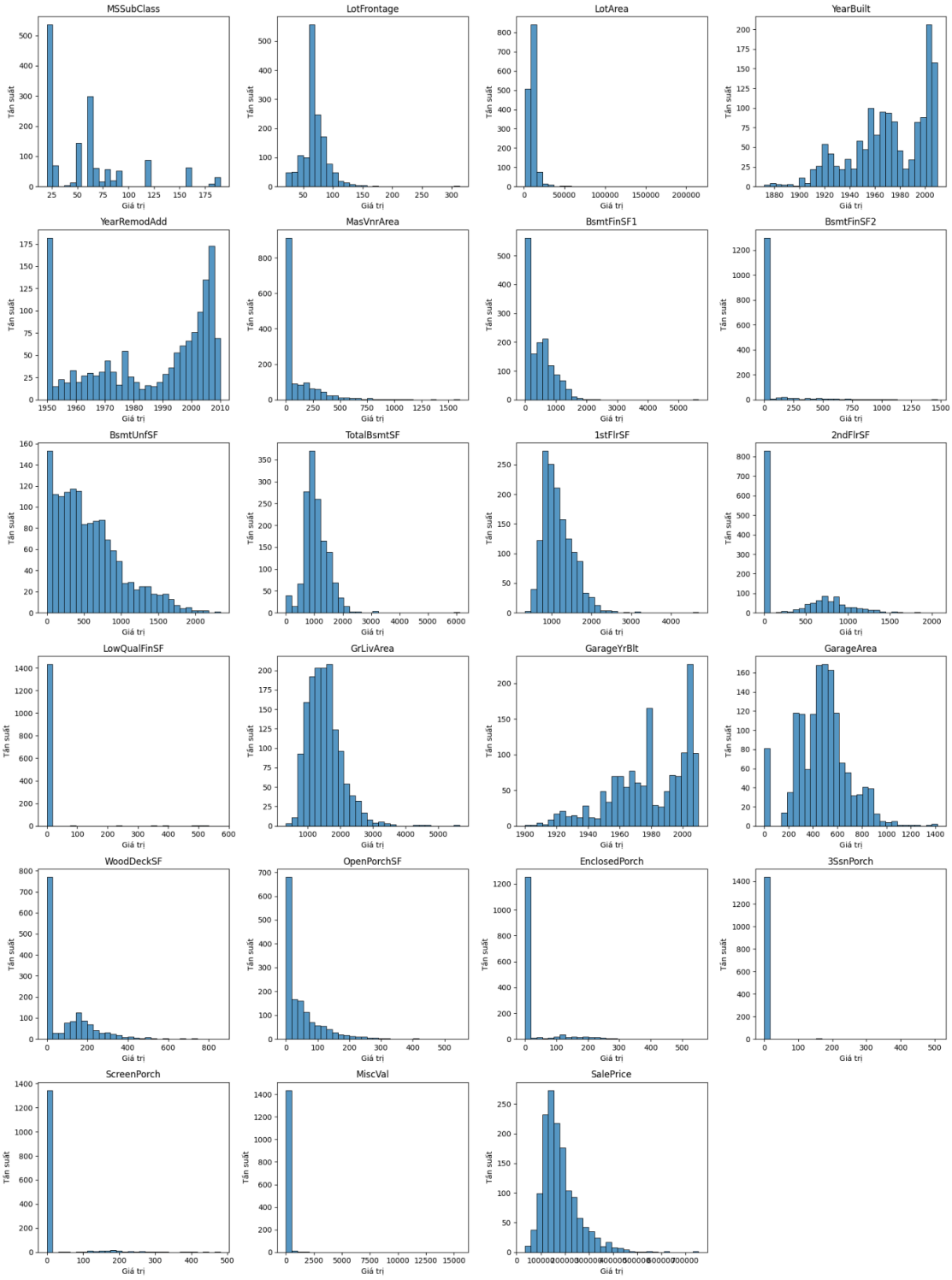
- Thực hiện xử lý outlier đối với các cột dữ liệu có dạng số và có nhiều giá trị khác nhau vì các cột này thường chứa thông tin phong phú hơn về phân phối dữ liệu, và outlier có thể ảnh hưởng lớn đến các phân tích thống kê hoặc mô hình học máy. Trong những cột số với nhiều giá trị, outlier có thể làm sai lệch các chỉ số thống kê như trung bình, độ lệch chuẩn và tương quan, dẫn đến các quyết định không chính xác. Ngược lại, việc loại bỏ hoặc xử lý outlier trong các cột có ít giá trị sẽ có nguy cơ mất thông tin

quan trọng. Nếu một giá trị hiếm có ý nghĩa cụ thể, việc loại bỏ nó có thể làm giảm độ chính xác của mô hình.



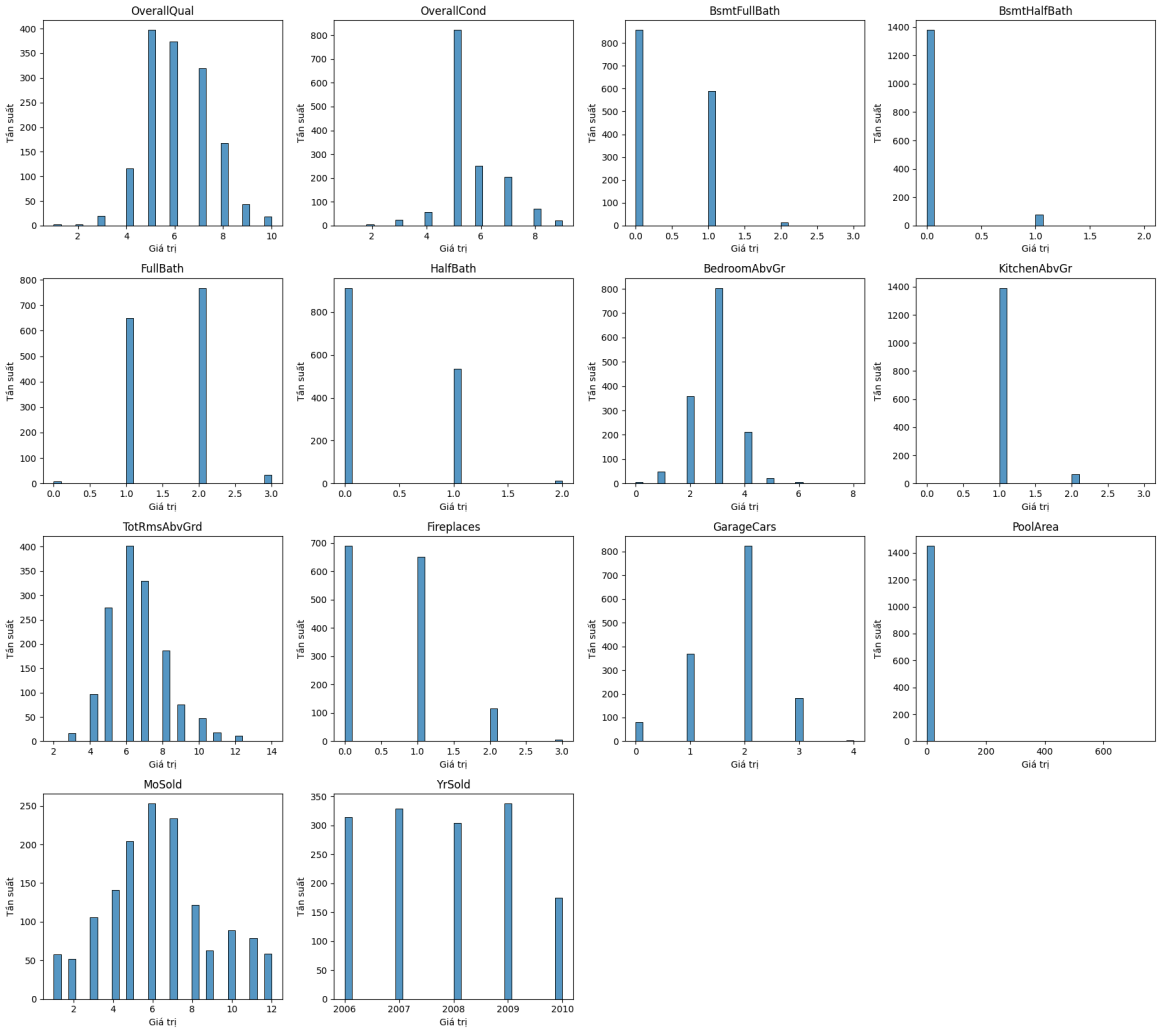
- Nhóm chia số lượng giá trị mỗi cột thành 2 tập: nhỏ hơn hoặc bằng 12, và lớn hơn 12 để dễ xử lý.

```
Số lượng cột có nhiều giá trị duy nhất: 23
Số lượng cột có ít giá trị duy nhất: 14
```

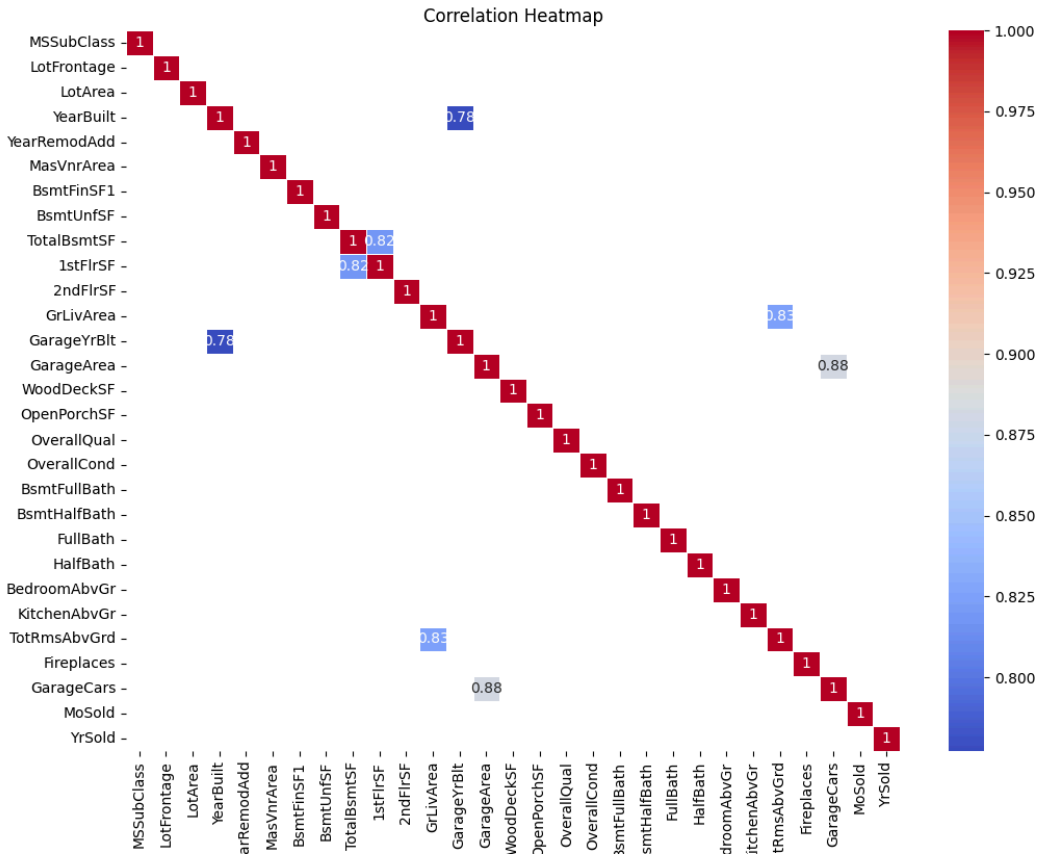


- Sau khi in ra biểu đồ phân phối của các cột có nhiều hơn 12 giá trị, nhận thấy các cột 'ScreenPorch', 'LowQualFinSF', '3SsnPorch', 'BsmtFinSF2', 'MiscVal', 'EnclosedPorch' có độ lệch cao vì các cột này có thể chứa nhiều giá trị ngoại lệ (outlier), làm sai lệch các dự đoán và kết quả của mô hình.

Điều này có thể dẫn đến việc mô hình không phản ánh đúng mối quan hệ giữa các biến nên remove 6 cột này.



- Tiếp theo, xem xét các cột có không quá 12 giá trị, nhóm thấy cột ‘PoolArea’ cũng có độ lệch cao, nên remove cột này.



- Sau khi loại bỏ bớt một số cột, tính ma trận tương quan đối với các biến số còn lại để xem xét mức độ tương quan giữa các cột với nhau:

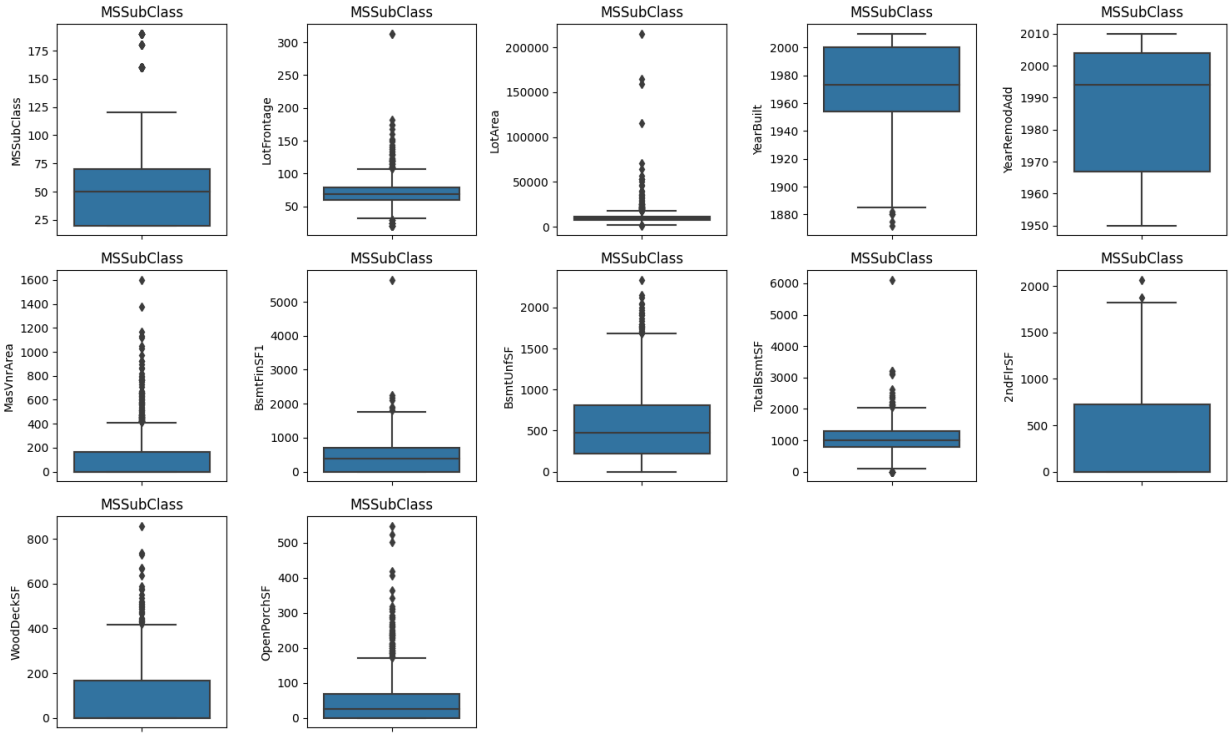
```
Cặp giá trị tương quan cao:
1stFlrSF and TotalBsmtSF: 0.82
GarageYrBlt and YearBuilt: 0.78
TotRmsAbvGrd and GrLivArea: 0.83
GarageCars and GarageArea: 0.88
```

- Thấy có 4 cặp cột số có mức độ tương quan lớn hơn 0.75%, loại bỏ mỗi cặp 1 cột để giảm bớt tình trạng đa cộng tuyến và giúp đơn giản hóa mô hình.

	MSSubClass	LotFrontage	LotArea	YearBuilt	YearRemodAdd	MasVnrArea	BsmtFinSF1	BsmtUnfSF
0	60	65.0	8450	2003	2003	196.0	706	150
1	20	80.0	9600	1976	1976	0.0	978	284
2	60	68.0	11250	2001	2002	162.0	486	434
3	70	60.0	9550	1915	1970	0.0	216	540
4	60	84.0	14260	2000	2000	350.0	655	490
...	...	...	...	...	...	...	...	...
1455	60	62.0	7917	1999	2000	0.0	0	953
1456	20	85.0	13175	1978	1988	119.0	790	589
1457	70	66.0	9042	1941	2006	0.0	275	877
1458	20	68.0	9717	1950	1996	0.0	49	0
1459	20	75.0	9937	1965	1965	0.0	830	136

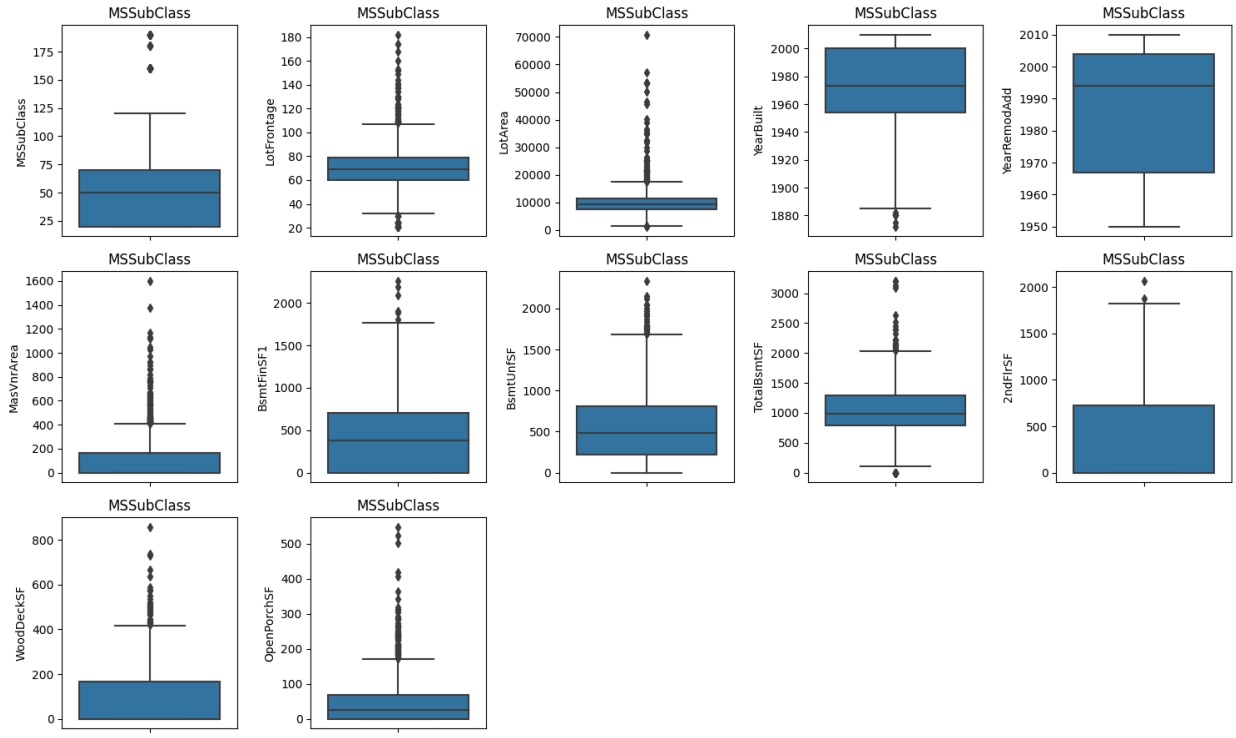
1460 rows × 25 columns

- Sau khi thực hiện bỏ bớt một số cột, số cột kiểu số còn lại của tập dữ liệu là 25. Tiếp theo, in ra boxplot cho các cột có nhiều giá trị duy nhất để xem xét các outlier:



- Từ các boxplot cho thấy, có 5 cột cần xử lý outlier: *LotFrontage, LotArea, BsmtFinSF1, TotalBsmtSF, TotRmsAbvGrd.* Và số lượng outlier trong mỗi

cột là ít, nên nhóm thực hiện drop các dòng có chứa outlier này. Mức ngưỡng lần lượt cho từng cột là: 200, 100000, 4000, 4000. 4000.



3. Encode dữ liệu:

- Chia các biến phân loại thành 2 nhóm: cột ordinal và nominal. Trong đó, ordinal là các biến phân loại có thứ tự, tức là các giá trị có thể được sắp xếp theo một thứ tự nhất định (ví dụ: chất lượng từ kém đến tốt). Còn nominal là các biến phân loại không có thứ tự, tức là các giá trị chỉ đại diện cho các loại khác nhau mà không có mối quan hệ thứ tự (ví dụ: kiểu nhà, khu phố).
    + **Ordinal columns**: ['ExterQual', 'ExterCond', 'BsmtQual', 'BsmtCond', 'HeatingQC', 'KitchenQual', 'FireplaceQu', 'GarageQual', 'GarageCond', 'PoolQC', 'BsmtExposure', 'Functional', 'LandSlope', 'PavedDrive', 'GarageFinish', 'BsmtFinType1', 'BsmtFinType2']. Đây là các cột chứa thông tin về chất lượng hoặc của các đặc điểm mà ngôi nhà có ví dụ như Chất lượng nhà bếp, Chất lượng ga-ra, Điều kiện của lò sưởi,vvv. Giá trị một dòng có thể có là:

```
Ex       Excellent
Gd       Good
TA       Average/Typical
Fa       Fair
Po       Poor
```

  + **Nominal columns:** ['MSZoning', 'Street', 'LotShape', 'LandContour', 'Utilities', 'LotConfig', 'Neighborhood', 'Condition1', 'Condition2', 'BldgType', 'HouseStyle', 'RoofStyle', 'RoofMatl', 'Exterior1st', 'Exterior2nd', 'MasVnrType', 'Foundation', 'Heating', 'CentralAir', 'Electrical', 'GarageType', 'SaleType', 'SaleCondition']. Các giá trị của những cột này mang thông tin đại diện cho các loại khác nhau mà không có tính thứ tự, ví dụ như các thông tin về Đường sá, Hàng xóm,vvv

- Đối với các cột **Ordinal**, sử dụng **LabelEncoder** để chuyển đổi các giá trị trong các cột ordinal thành các số nguyên. Mỗi giá trị duy nhất trong cột sẽ được gán một số nguyên khác nhau. Điều này giúp mô hình nhận biết được thứ tự giữa các giá trị.

	MSSubClass	LotFrontage	LotArea	YearBuilt	YearRemodAdd	MasVnrArea	BsmtFinSF1	BsmtUnfSF
0	60	65.0	8450	2003	2003	196.0	706	150
1	20	80.0	9600	1976	1976	0.0	978	284
2	60	68.0	11250	2001	2002	162.0	486	434
3	70	60.0	9550	1915	1970	0.0	216	540
4	60	84.0	14260	2000	2000	350.0	655	490
...	...	...	...	...	...	...	...	...
1455	60	62.0	7917	1999	2000	0.0	0	953
1456	20	85.0	13175	1978	1988	119.0	790	589
1457	70	66.0	9042	1941	2006	0.0	275	877
1458	20	68.0	9717	1950	1996	0.0	49	0
1459	20	75.0	9937	1965	1965	0.0	830	136

1454 rows × 64 columns

- Các cột Nominal thì dùng **One-Hot Encoding** để tạo các cột nhị phân bằng hàm get_dummies.

```
<class 'pandas.core.frame.DataFrame'>
Index: 1454 entries, 0 to 1459
Columns: 222 entries, MSSubClass to SaleCondition_Partial
dtypes: bool(184), float64(2), int64(36)
memory usage: 704.3 KB
```

Sau khi thực hiện Encode với các cột dạng phân loại, tập dữ liệu có tổng cộng 222 cột, trong đó, có 184 cột kiểu bool, 2 cột kiểu float và 36 cột kiểu int.

4. Chuẩn hóa dữ liệu:

- Chiến thuật chuẩn hóa dữ liệu nhóm sử dụng là **Standardization** (chuẩn hóa theo chuẩn), được thực hiện thông qua **StandardScaler** từ thư viện sklearn.
- Chỉ thực hiện chuẩn hóa đối với các cột thuộc dạng số ban đầu (trước khi thực hiện Encode)
- Mục tiêu của chuẩn hóa là biến đổi dữ liệu sao cho nó có trung bình bằng 0 và độ lệch chuẩn bằng 1

```
df[numerical_cols]
```

	MSSubClass	LotFrontage	LotArea	YearBuilt	YearRemodAdd	MasVnrArea	BsmtFinSF1	BsmtUnfSF
0	0.073836	-0.223516	-0.292989	1.049501	0.880128	0.520741	0.616203	-0.945995
1	-0.874133	0.525544	-0.082066	0.156499	-0.428117	-0.570040	1.242232	-0.642892
2	0.073836	-0.073704	0.220563	0.983353	0.831675	0.331524	0.109855	-0.303599
3	0.310828	-0.473203	-0.091236	-1.861023	-0.718838	-0.570040	-0.511571	-0.063831
4	0.073836	0.725293	0.772632	0.950278	0.734768	1.377783	0.498822	-0.176929
...	...	...	...	...	...	...	...	...
1455	0.073836	-0.373328	-0.390747	0.917204	0.734768	-0.570040	-1.008712	0.870357
1456	-0.874133	0.775231	0.573630	0.222648	0.153325	0.092220	0.809535	0.047005
1457	0.310828	-0.173579	-0.184409	-1.001095	1.025489	-0.570040	-0.375778	0.698449
1458	-0.874133	-0.073704	-0.060607	-0.703428	0.540953	-0.570040	-0.895935	-1.285288
1459	-0.874133	0.275857	-0.020256	-0.207316	-0.961106	-0.570040	0.901598	-0.977662

1454 rows × 38 columns

.Kết quả là các giá trị trong cột sẽ nằm trong khoảng gần xung quanh 0, giúp các thuật toán học máy hoạt động hiệu quả hơn.