Exploratory Data Analysis

In this section I simple check variables and their properties. Try to get insight will data enough to train usefull model.

You can find descriptions of the data from README.md.

Notebook Structure

- **Describing Data**
- **Correlation**
    - **The most Correlated 10 Features with SalePrice**
- **Demonstrating Categorical Data Unique Values**

In [6]:
```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from pathlib import Path
from src.utils import html_table

raw_data_dir = Path('../data/raw/')
processed_data_dir = Path('../data/processed/')
file_name = 'train.csv'
file_path = raw_data_dir / file_name
```

Describing Data

In [2]:
```python
df = pd.read_csv(file_path)
df.head()
```

Out[2]:

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	PoolArea	PoolQC	F
0	1	60	RL	65.0	8450	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	
1	2	20	RL	80.0	9600	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	
2	3	60	RL	68.0	11250	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	
3	4	70	RL	60.0	9550	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	
4	5	60	RL	84.0	14260	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	

5 rows × 81 columns

In [3]:
```python
df.drop(axis=1, columns=['Id'], inplace=True)
df.shape
```

Out[3]: (1460, 80)

In [4]:
```python
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1460 entries, 0 to 1459
Data columns (total 80 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   MSSubClass     1460 non-null   int64  
 1   MSZoning       1460 non-null   object 
 2   LotFrontage    1201 non-null   float64
 3   LotArea        1460 non-null   int64  
 4   Street         1460 non-null   object 
 5   Alley          91 non-null     object 
 6   LotShape       1460 non-null   object 
 7   LandContour    1460 non-null   object 
 8   Utilities      1460 non-null   object 
 9   LotConfig      1460 non-null   object 
 10  LandSlope      1460 non-null   object 
 11  Neighborhood   1460 non-null   object 
 12  Condition1     1460 non-null   object 
 13  Condition2     1460 non-null   object 
 14  BldgType       1460 non-null   object 
 15  HouseStyle     1460 non-null   object 
 16  OverallQual    1460 non-null   int64  
 17  OverallCond    1460 non-null   int64  
 18  YearBuilt      1460 non-null   int64  
 19  YearRemodAdd   1460 non-null   int64  
 20  RoofStyle      1460 non-null   object 
 21  RoofMatl       1460 non-null   object 
 22  Exterior1st    1460 non-null   object 
 23  Exterior2nd    1460 non-null   object 
 24  MasVnrType     1452 non-null   object 
 25  MasVnrArea     1452 non-null   float64
 26  ExterQual      1460 non-null   object 
 27  ExterCond      1460 non-null   object 
 28  Foundation     1460 non-null   object 
 29  BsmtQual       1423 non-null   object 
 30  BsmtCond       1423 non-null   object 
 31  BsmtExposure   1422 non-null   object 
 32  BsmtFinType1   1423 non-null   object 
 33  BsmtFinSF1     1460 non-null   int64  
 34  BsmtFinType2   1422 non-null   object 
 35  BsmtFinSF2     1460 non-null   int64  
 36  BsmtUnfSF      1460 non-null   int64  
 37  TotalBsmtSF    1460 non-null   int64  
 38  Heating        1460 non-null   object 
 39  HeatingQC      1460 non-null   object 
 40  CentralAir     1460 non-null   object 
 41  Electrical     1459 non-null   object 
 42  1stFlrSF       1460 non-null   int64  
 43  2ndFlrSF       1460 non-null   int64  
 44  LowQualFinSF   1460 non-null   int64  
 45  GrLivArea      1460 non-null   int64  
 46  BsmtFullBath   1460 non-null   int64  
 47  BsmtHalfBath   1460 non-null   int64  
 48  FullBath       1460 non-null   int64  
 49  HalfBath       1460 non-null   int64  
 50  BedroomAbvGr   1460 non-null   int64  
 51  KitchenAbvGr   1460 non-null   int64  
 52  KitchenQual    1460 non-null   object 
 53  TotRmsAbvGrd   1460 non-null   int64  
 54  Functional     1460 non-null   object 
 55  Fireplaces     1460 non-null   int64  
 56  FireplaceQu    770 non-null    object 
 57  GarageType     1379 non-null   object 
 58  GarageYrBlt    1379 non-null   float64
 59  GarageFinish   1379 non-null   object 
 60  GarageCars     1460 non-null   int64  
 61  GarageArea     1460 non-null   int64  
 62  GarageQual     1379 non-null   object 
 63  GarageCond     1379 non-null   object 
 64  PavedDrive     1460 non-null   object 
 65  WoodDeckSF     1460 non-null   int64  
 66  OpenPorchSF    1460 non-null   int64  
 67  EnclosedPorch  1460 non-null   int64  
 68  3SsnPorch      1460 non-null   int64  
 69  ScreenPorch    1460 non-null   int64  
 70  PoolArea       1460 non-null   int64  
 71  PoolQC         7 non-null      object 
 72  Fence          281 non-null    object 
 73  MiscFeature    54 non-null     object 
 74  MiscVal        1460 non-null   int64  
 75  MoSold         1460 non-null   int64  
 76  YrSold         1460 non-null   int64  
 77  SaleType       1460 non-null   object 
 78  SaleCondition  1460 non-null   object 
 79  SalePrice      1460 non-null   int64  
dtypes: float64(3), int64(34), object(43)
memory usage: 912.6+ KB
```

In [7]:
```python
html_table(df.describe().T)
```

	count	mean	std	min	25%	50%	75%	max
MSSubClass	1460.0	56.897260	42.300571	20.0	20.00	50.0	70.00	190.0
LotFrontage	1201.0	70.049958	24.284752	21.0	59.00	69.0	80.00	313.0
LotArea	1460.0	10516.828082	9981.264932	1300.0	7553.50	9478.5	11601.50	215245.0
OverallQual	1460.0	6.099315	1.382997	1.0	5.00	6.0	7.00	10.0
OverallCond	1460.0	5.575342	1.112799	1.0	5.00	5.0	6.00	9.0
YearBuilt	1460.0	1971.267808	30.202904	1872.0	1954.00	1973.0	2000.00	2010.0
YearRemodAdd	1460.0	1984.865753	20.645407	1950.0	1967.00	1994.0	2004.00	2010.0
MasVnrArea	1452.0	103.685262	181.066207	0.0	0.00	0.0	166.00	1600.0
BsmtFinSF1	1460.0	443.639726	456.098091	0.0	0.00	383.5	712.25	5644.0
BsmtFinSF2	1460.0	46.549315	161.319273	0.0	0.00	0.0	0.00	1474.0
BsmtUnfSF	1460.0	567.240411	441.866955	0.0	223.00	477.5	808.00	2336.0
TotalBsmtSF	1460.0	1057.429452	438.705324	0.0	795.75	991.5	1298.25	6110.0
1stFlrSF	1460.0	1162.626712	386.587738	334.0	882.00	1087.0	1391.25	4692.0
2ndFlrSF	1460.0	346.992466	436.528436	0.0	0.00	0.0	728.00	2065.0
LowQualFinSF	1460.0	5.844521	48.623081	0.0	0.00	0.0	0.00	572.0
GrLivArea	1460.0	1515.463699	525.480383	334.0	1129.50	1464.0	1776.75	5642.0
BsmtFullBath	1460.0	0.425342	0.518911	0.0	0.00	0.0	1.00	3.0
BsmtHalfBath	1460.0	0.057534	0.238753	0.0	0.00	0.0	0.00	2.0
FullBath	1460.0	1.565068	0.550916	0.0	1.00	2.0	2.00	3.0
HalfBath	1460.0	0.382877	0.502885	0.0	0.00	0.0	1.00	2.0
BedroomAbvGr	1460.0	2.866438	0.815778	0.0	2.00	3.0	3.00	8.0
KitchenAbvGr	1460.0	1.046575	0.220338	0.0	1.00	1.0	1.00	3.0
TotRmsAbvGrd	1460.0	6.517808	1.625393	2.0	5.00	6.0	7.00	14.0
Fireplaces	1460.0	0.613014	0.644666	0.0	0.00	1.0	1.00	3.0
GarageYrBlt	1379.0	1978.506164	24.689725	1900.0	1961.00	1980.0	2002.00	2010.0
GarageCars	1460.0	1.767123	0.747315	0.0	1.00	2.0	2.00	4.0
GarageArea	1460.0	472.980137	213.804841	0.0	334.50	480.0	576.00	1418.0
WoodDeckSF	1460.0	94.244521	125.338794	0.0	0.00	0.0	168.00	857.0
OpenPorchSF	1460.0	46.660274	66.256028	0.0	0.00	25.0	68.00	547.0
EnclosedPorch	1460.0	21.954110	61.119149	0.0	0.00	0.0	0.00	552.0
3SsnPorch	1460.0	3.409589	29.317331	0.0	0.00	0.0	0.00	508.0
ScreenPorch	1460.0	15.060959	55.757415	0.0	0.00	0.0	0.00	480.0
PoolArea	1460.0	2.758904	40.177307	0.0	0.00	0.0	0.00	738.0
MiscVal	1460.0	43.489041	496.123024	0.0	0.00	0.0	0.00	15500.0
MoSold	1460.0	6.321918	2.703626	1.0	5.00	6.0	8.00	12.0
YrSold	1460.0	2007.815753	1.328095	2006.0	2007.00	2008.0	2009.00	2010.0
SalePrice	1460.0	180921.195890	79442.502883	34900.0	129975.00	163000.0	214000.00	755000.0

Correlation

I looked for a general correlation between continues variables and after that because of our main goal of predicting SalePrices I focused the correlation on SalePricing I found that there are lots of variables that are highly correlated with SalePrice it seems like we can fit a good model. You can see the results below.

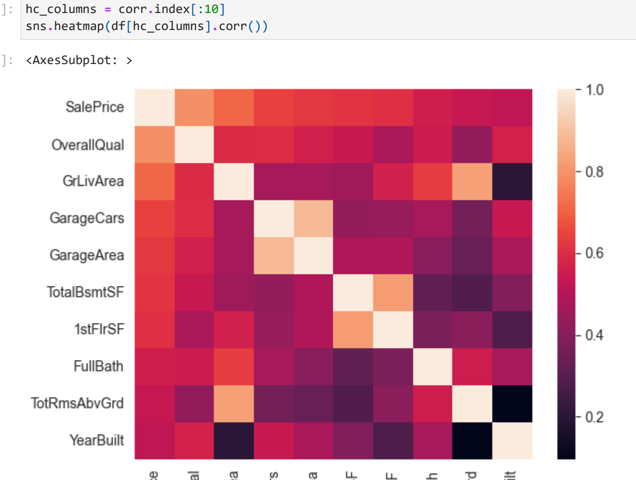
In [9]:
```python
corr = df.corr(numeric_only=True).sort_values(by='SalePrice', ascending=False,
                                              key=lambda val: abs(val))
html_table(corr)
```

	MSSubClass	LotFrontage	LotArea	OverallQual	OverallCond	YearBuilt	YearRemodAdd	MasVnrArea	BsmtFi
SalePrice	-0.084284	0.351799	0.263843	0.790982	-0.077856	0.522897	0.507101	0.477493	0.38
OverallQual	0.032628	0.251646	0.105806	1.000000	-0.091932	0.572323	0.550684	0.411876	0.23
GrLivArea	0.074853	0.402797	0.263116	0.593007	-0.079686	0.199010	0.287389	0.390857	0.20
GarageCars	-0.040110	0.285691	0.154871	0.600671	-0.185758	0.537850	0.420622	0.364204	0.22
GarageArea	-0.098672	0.344997	0.180403	0.562022	-0.151521	0.478954	0.371600	0.373066	0.29
TotalBsmtSF	-0.238518	0.392075	0.260833	0.537808	-0.171098	0.391452	0.291066	0.363936	0.52
1stFlrSF	-0.251758	0.457181	0.299475	0.476224	-0.144203	0.281986	0.240379	0.344501	0.44
FullBath	0.131608	0.198769	0.126031	0.550600	-0.194149	0.468271	0.439046	0.276833	0.05
TotRmsAbvGrd	0.040380	0.352096	0.190015	0.427452	-0.057583	0.095589	0.191740	0.280682	0.04
YearBuilt	0.027850	0.123349	0.014228	0.572323	-0.375983	1.000000	0.592855	0.315707	0.24
YearRemodAdd	0.040581	0.088866	0.013788	0.550684	0.073741	0.592855	1.000000	0.179618	0.12
GarageYrBlt	0.085072	0.070250	-0.024947	0.547766	-0.324297	0.825667	0.642277	0.252691	0.15
MasVnrArea	0.022936	0.193458	0.104160	0.411876	-0.128101	0.315707	0.179618	1.000000	0.26
Fireplaces	-0.045569	0.266639	0.271364	0.396765	-0.023820	0.147716	0.112581	0.249070	0.26
BsmtFinSF1	-0.069836	0.233633	0.214103	0.239666	-0.046231	0.249503	0.128451	0.264736	1.00
LotFrontage	-0.386347	1.000000	0.426095	0.251646	-0.059213	0.123349	0.088866	0.193458	0.23
WoodDeckSF	-0.012579	0.088521	0.171698	0.238923	-0.003334	0.224880	0.205726	0.159718	0.20
2ndFlrSF	0.307886	0.080177	0.050986	0.295493	0.028942	0.010308	0.140024	0.174561	-0.13
OpenPorchSF	-0.006100	0.151972	0.084774	0.308819	-0.032589	0.188686	0.226298	0.125703	0.11
HalfBath	0.177354	0.053532	0.014259	0.273458	-0.060769	0.242656	0.183331	0.201444	0.00
LotArea	-0.139781	0.426095	1.000000	0.105806	-0.005636	0.014228	0.013788	0.104160	0.21
BsmtFullBath	0.003491	0.100949	0.158155	0.111098	-0.054942	0.187599	0.119470	0.085310	0.64
BsmtUnfSF	-0.140759	0.132644	-0.002618	0.308159	-0.136841	0.149040	0.181133	0.114442	-0.49
BedroomAbvGr	-0.023438	0.263170	0.119690	0.101676	0.012980	-0.070651	-0.040581	0.102821	-0.10
KitchenAbvGr	0.281721	-0.006069	-0.017784	-0.183882	-0.087001	-0.174800	-0.149598	-0.037610	-0.08
EnclosedPorch	-0.012037	0.010700	-0.018340	-0.113937	0.070356	-0.387268	-0.193919	-0.110204	-0.10
ScreenPorch	-0.026030	0.041383	0.043160	0.064886	0.054811	-0.050364	-0.038740	0.061466	0.06
PoolArea	0.008283	0.206167	0.077672	0.065166	-0.001985	0.004950	0.005829	0.011723	0.14
MSSubClass	1.000000	-0.386347	-0.139781	0.032628	-0.059316	0.027850	0.040581	0.022936	-0.06
OverallCond	-0.059316	-0.059213	-0.005636	-0.091932	1.000000	-0.375983	0.073741	-0.128101	-0.04
MoSold	-0.013585	0.011200	0.001205	0.070815	-0.003511	0.012398	0.021490	-0.005965	-0.01
3SsnPorch	-0.043825	0.070029	0.020423	0.030371	0.025504	0.031355	0.045286	0.018796	0.02
YrSold	-0.021407	0.007450	-0.014261	-0.027347	0.043950	-0.013618	0.035743	-0.008201	0.01
LowQualFinSF	0.046474	0.038469	0.004779	-0.030429	0.025494	-0.183784	-0.062419	-0.069071	-0.06
MiscVal	-0.007683	0.003368	0.038068	-0.031406	0.068777	-0.034383	-0.010286	-0.029815	0.00
BsmtHalfBath	-0.002333	-0.007234	0.048046	-0.040150	0.117821	-0.038162	-0.012337	0.026673	0.06
BsmtFinSF2	-0.065649	0.049900	0.111170	-0.059119	0.040229	-0.049107	-0.067759	-0.072319	-0.05

The most Correlated 10 Features with SalePrice

In [10]:
```python
hc_columns = corr.index[:10]
sns.heatmap(df[hc_columns].corr())
```

Out[10]: <AxesSubplot: >



In [11]:
```python
df[hc_columns].info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1460 entries, 0 to 1459
Data columns (total 10 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   SalePrice     1460 non-null   int64
 1   OverallQual   1460 non-null   int64
 2   GrLivArea     1460 non-null   int64
 3   GarageCars    1460 non-null   int64
 4   GarageArea    1460 non-null   int64
 5   TotalBsmtSF   1460 non-null   int64
 6   1stFlrSF      1460 non-null   int64
 7   FullBath      1460 non-null   int64
 8   TotRmsAbvGrd  1460 non-null   int64
 9   YearBuilt     1460 non-null   int64
dtypes: int64(10)
memory usage: 114.2 KB
```

In [16]:
```python
rows = 5
cols = 2
fig, axs = plt.subplots(rows, cols, figsize=[9, 18], dpi=150)
fig.tight_layout(pad=4.0)
for c in range(2):
    for r in range(5):
        sns.scatterplot(x=df[corr.index[rows * c + r + 1]], y=df['SalePrice'],
                        ax=axs[r, c])
```

Demonstrating Categorical Data Unique Values

In [17]:
```python
isCategorical = (df.dtypes == object)
cat_cols = isCategorical[isCategorical > 0].index
```

In [18]:
```python
unique_count = df[cat_cols].nunique().sort_values(ascending=False)
unique_count
```

Out[18]:
```
Neighborhood     25
Exterior2nd      16
Exterior1st      15
SaleType          9
Condition1        9
Condition2        8
HouseStyle        8
RoofMatl          8
Functional        7
BsmtFinType2      6
Heating           6
RoofStyle         6
SaleCondition     6
BsmtFinType1      6
GarageType        6
Foundation        6
Electrical        5
FireplaceQu       5
HeatingQC         5
GarageQual        5
GarageCond        5
MSZoning          5
LotConfig         5
ExterCond         5
BldgType          5
BsmtExposure      4
MiscFeature       4
Fence             4
LotShape          4
LandContour       4
BsmtCond          4
KitchenQual       4
MasVnrType        4
ExterQual         4
BsmtQual          4
LandSlope         3
GarageFinish      3
PavedDrive        3
PoolQC            3
Utilities         2
CentralAir        2
Street            2
Alley             2
dtype: int64
```

In [20]:
```python
# Demonstrate all categories
max_cat = unique_count[0]
categories = pd.DataFrame(columns=range(max_cat))
for col_name in list(cat_cols):
    unqs = df[col_name].unique()
    if len(unqs) == max_cat:
        categories.loc[col_name] = unqs
        continue
    categories.loc[col_name] = np.concatenate(
        (unqs, np.array([np.nan, ] * (max_cat - len(unqs)))))

html_table(categories)
```

	0	1	2	3	4	5	6	7	8	9	10
MSZoning	RL	RM	C (all)	FV	RH	NaN	NaN	NaN	NaN	NaN	NaN
Street	Pave	Grvl	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Alley	NaN	Grvl	Pave	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
LotShape	Reg	IR1	IR2	IR3	NaN	NaN	NaN	NaN	NaN	NaN	NaN
LandContour	Lvl	Bnk	Low	HLS	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Utilities	AllPub	NoSeWa	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
LotConfig	Inside	FR2	Corner	CulDSac	FR3	NaN	NaN	NaN	NaN	NaN	NaN
LandSlope	Gtl	Mod	Sev	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Neighborhood	CollgCr	Veenker	Crawfor	NoRidge	Mitchel	Somerst	NWAmes	OldTown	BrkSide	Sawyer	NridgHt
Condition1	Norm	Feedr	PosN	Artery	RRAe	RRNn	RRAn	PosA	RRNe	NaN	NaN
Condition2	Norm	Artery	RRNn	Feedr	PosN	PosA	RRAn	RRAe	NaN	NaN	NaN
BldgType	1Fam	2fmCon	Duplex	TwnhsE	Twnhs	NaN	NaN	NaN	NaN	NaN	NaN
HouseStyle	2Story	1Story	1.5Fin	1.5Unf	SFoyer	SLvl	2.5Unf	2.5Fin	NaN	NaN	NaN
RoofStyle	Gable	Hip	Gambrel	Mansard	Flat	Shed	NaN	NaN	NaN	NaN	NaN
RoofMatl	CompShg	WdShngl	Metal	WdShake	Membran	Tar&Grv	Roll	ClyTile	NaN	NaN	NaN
Exterior1st	VinylSd	MetalSd	Wd Sdng	HdBoard	BrkFace	WdShing	CemntBd	Plywood	AsbShng	Stucco	BrkComm
Exterior2nd	VinylSd	MetalSd	Wd Shng	HdBoard	Plywood	Wd Sdng	CmentBd	BrkFace	Stucco	AsbShng	Brk Cmn
MasVnrType	BrkFace	None	Stone	BrkCmn	NaN	NaN	NaN	NaN	NaN	NaN	NaN
ExterQual	Gd	TA	Ex	Fa	NaN	NaN	NaN	NaN	NaN	NaN	NaN
ExterCond	TA	Gd	Fa	Po	Ex	NaN	NaN	NaN	NaN	NaN	NaN
Foundation	PConc	CBlock	BrkTil	Wood	Slab	Stone	NaN	NaN	NaN	NaN	NaN
BsmtQual	Gd	TA	Ex	NaN	Fa	NaN	NaN	NaN	NaN	NaN	NaN
BsmtCond	TA	Gd	NaN	Fa	Po	NaN	NaN	NaN	NaN	NaN	NaN
BsmtExposure	No	Gd	Mn	Av	NaN	NaN	NaN	NaN	NaN	NaN	NaN
BsmtFinType1	GLQ	ALQ	Unf	Rec	BLQ	NaN	LwQ	NaN	NaN	NaN	NaN
BsmtFinType2	Unf	BLQ	NaN	ALQ	Rec	LwQ	GLQ	NaN	NaN	NaN	NaN
Heating	GasA	GasW	Grav	Wall	OthW	Floor	NaN	NaN	NaN	NaN	NaN
HeatingQC	Ex	Gd	TA	Fa	Po	NaN	NaN	NaN	NaN	NaN	NaN
CentralAir	Y	N	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Electrical	SBrkr	FuseF	FuseA	FuseP	Mix	NaN	NaN	NaN	NaN	NaN	NaN
KitchenQual	Gd	TA	Ex	Fa	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Functional	Typ	Min1	Maj1	Min2	Mod	Maj2	Sev	NaN	NaN	NaN	NaN
FireplaceQu	NaN	TA	Gd	Fa	Ex	Po	NaN	NaN	NaN	NaN	NaN
GarageType	Attchd	Detchd	BuiltIn	CarPort	NaN	Basment	2Types	NaN	NaN	NaN	NaN
GarageFinish	RFn	Unf	Fin	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
GarageQual	TA	Fa	Gd	NaN	Ex	Po	NaN	NaN	NaN	NaN	NaN
GarageCond	TA	Fa	NaN	Gd	Po	Ex	NaN	NaN	NaN	NaN	NaN
PavedDrive	Y	N	P	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
PoolQC	NaN	Ex	Fa	Gd	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Fence	NaN	MnPrv	GdWo	GdPrv	MnWw	NaN	NaN	NaN	NaN	NaN	NaN
MiscFeature	NaN	Shed	Gar2	Othr	TenC	NaN	NaN	NaN	NaN	NaN	NaN
SaleType	WD	New	COD	ConLD	ConLI	CWD	ConLw	Con	Oth	NaN	NaN
SaleCondition	Normal	Abnorml	Partial	AdjLand	Alloca	Family	NaN	NaN	NaN	NaN	NaN