krishnaik06 / Advanced-House-Price-Prediction-

Code | Issues 2 | Pull requests 2 | Actions | Projects | Wiki | Security | Insights

master | Advanced-House-Price-Prediction- / Feature Engineering.ipynb | Go to file | ···

krishnaik06 Add files via upload | Latest commit b8f8c3b on Feb 6, 2020 | History

1 contributor

17983 lines (17983 sloc) | 663 KB | Raw | Blame

Advanced Housing Prices- Feature Engineering

The main aim of this project is to predict the house price based on various features which we will discuss as we go ahead

Dataset to downloaded from the below link

https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data (https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data)

We will be performing all the below steps in Feature Engineering

1. Missing values
2. Temporal variables
3. Categorical variables: remove rare labels
4. Standarise the values of the variables to the same range

In [138]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
# to visualise al the columns in the dataframe
pd.pandas.set_option('display.max_columns', None)
```

In [139]:

```
dataset=pd.read_csv('train.csv')
dataset.head()
```

Out[139]:

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	Neighborhood	Condition1	Cor
0	1	60	RL	65.0	8450	Pave	NaN	Reg	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Nor
1	2	20	RL	80.0	9600	Pave	NaN	Reg	Lvl	AllPub	FR2	Gtl	Veenker	Feedr	Nor
2	3	60	RL	68.0	11250	Pave	NaN	IR1	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Nor
3	4	70	RL	60.0	9550	Pave	NaN	IR1	Lvl	AllPub	Corner	Gtl	Crawfor	Norm	Nor
4	5	60	RL	84.0	14260	Pave	NaN	IR1	Lvl	AllPub	FR2	Gtl	NoRidge	Norm	Nor

In [140]:

```
## Always remember there way always be a chance of data leakage so we need to split the data first and then apply feature
## Engineering
from sklearn.model_selection import train_test_split
X_train,X_test,y_train,y_test=train_test_split(dataset,dataset['SalePrice'],test_size=0.1,random_state=0)
```

In [141]:
```
X_train.shape, X_test.shape
```

Out[141]:
```
((1314, 81), (146, 81))
```

Missing Values

In [142]:
```
## Let us capture all the nan values
## First lets handle Categorical features which are missing
features_nan=[feature for feature in dataset.columns if dataset[feature].isnull().sum()>1 and dataset[feature].dtypes=='O']

for feature in features_nan:
    print("{}: {}% missing values".format(feature,np.round(dataset[feature].isnull().mean(),4)))
```

```
Alley: 0.9377% missing values
MasVnrType: 0.0055% missing values
BsmtQual: 0.0253% missing values
BsmtCond: 0.0253% missing values
BsmtExposure: 0.026% missing values
BsmtFinType1: 0.0253% missing values
BsmtFinType2: 0.026% missing values
FireplaceQu: 0.4726% missing values
GarageType: 0.0555% missing values
GarageFinish: 0.0555% missing values
GarageQual: 0.0555% missing values
GarageCond: 0.0555% missing values
PoolQC: 0.9952% missing values
Fence: 0.8075% missing values
MiscFeature: 0.963% missing values
```

In [143]:
```
## Replace missing value with a new label
def replace_cat_feature(dataset,features_nan):
    data=dataset.copy()
    data[features_nan]=data[features_nan].fillna('Missing')
    return data

dataset=replace_cat_feature(dataset,features_nan)

dataset[features_nan].isnull().sum()
```

Out[143]:
```
Alley           0
MasVnrType      0
BsmtQual        0
BsmtCond        0
BsmtExposure    0
BsmtFinType1    0
```

```
BsmtFinType2    0
FireplaceQu     0
GarageType      0
GarageFinish    0
GarageQual      0
GarageCond      0
PoolQC          0
Fence           0
MiscFeature     0
dtype: int64
```

In [144]:

```
dataset.head()
```

Out[144]:

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	Neighborhood	Condition1	C
0	1	60	RL	65.0	8450	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	N
1	2	20	RL	80.0	9600	Pave	Missing	Reg	Lvl	AllPub	FR2	Gtl	Veenker	Feedr	N
2	3	60	RL	68.0	11250	Pave	Missing	IR1	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	N
3	4	70	RL	60.0	9550	Pave	Missing	IR1	Lvl	AllPub	Corner	Gtl	Crawfor	Norm	N
4	5	60	RL	84.0	14260	Pave	Missing	IR1	Lvl	AllPub	FR2	Gtl	NoRidge	Norm	N

In [ ]:

In [145]:

```
## Now lets check for numerical variables the contains missing values
numerical_with_nan=[feature for feature in dataset.columns if dataset[feature].isnull().sum()>1 and dataset[feature].dtypes!='O']

## We will print the numerical nan variables and percentage of missing values

for feature in numerical_with_nan:
    print("{}: {}% missing value".format(feature,np.around(dataset[feature].isnull().mean(),4)))
```

```
LotFrontage: 0.1774% missing value
MasVnrArea: 0.0055% missing value
GarageYrBlt: 0.0555% missing value
```

In [ ]:

In [146]:

```
## Replacing the numerical Missing Values

for feature in numerical_with_nan:
    ## We will replace by using median since there are outliers
    median_value=dataset[feature].median()
    ## create a new feature to capture nan values
    dataset[feature+'nan']=np.where(dataset[feature].isnull(),1,0)
    dataset[feature].fillna(median_value,inplace=True)

dataset[numerical_with_nan].isnull().sum()
```

```
Out[146]: LotFrontage    0
          MasVnrArea     0
          GarageYrBlt    0
          dtype: int64
```

```
In [147]: dataset.head(50)
```

Out[147]:

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	Neighborhood	Condition1
0	1	60	RL	65.0	8450	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	CollgCr	Norm
1	2	20	RL	80.0	9600	Pave	Missing	Reg	Lvl	AllPub	FR2	Gtl	Veenker	Feedr
2	3	60	RL	68.0	11250	Pave	Missing	IR1	Lvl	AllPub	Inside	Gtl	CollgCr	Norm
3	4	70	RL	60.0	9550	Pave	Missing	IR1	Lvl	AllPub	Corner	Gtl	Crawfor	Norm
4	5	60	RL	84.0	14260	Pave	Missing	IR1	Lvl	AllPub	FR2	Gtl	NoRidge	Norm
5	6	50	RL	85.0	14115	Pave	Missing	IR1	Lvl	AllPub	Inside	Gtl	Mitchel	Norm
6	7	20	RL	75.0	10084	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	Somerst	Norm
7	8	60	RL	69.0	10382	Pave	Missing	IR1	Lvl	AllPub	Corner	Gtl	NWAmes	PosN
8	9	50	RM	51.0	6120	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	OldTown	Artery
9	10	190	RL	50.0	7420	Pave	Missing	Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Artery
10	11	20	RL	70.0	11200	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	Sawyer	Norm
11	12	60	RL	85.0	11924	Pave	Missing	IR1	Lvl	AllPub	Inside	Gtl	NridgHt	Norm
12	13	20	RL	69.0	12968	Pave	Missing	IR2	Lvl	AllPub	Inside	Gtl	Sawyer	Norm
13	14	20	RL	91.0	10652	Pave	Missing	IR1	Lvl	AllPub	Inside	Gtl	CollgCr	Norm
14	15	20	RL	69.0	10920	Pave	Missing	IR1	Lvl	AllPub	Corner	Gtl	NAmes	Norm
15	16	45	RM	51.0	6120	Pave	Missing	Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Norm
16	17	20	RL	69.0	11241	Pave	Missing	IR1	Lvl	AllPub	CulDSac	Gtl	NAmes	Norm
17	18	90	RL	72.0	10791	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	Sawyer	Norm
18	19	20	RL	66.0	13695	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	SawyerW	RRAe
19	20	20	RL	70.0	7560	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	NAmes	Norm

19	20	20	RL	70.0	7560	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	NAmes	Norm
20	21	60	RL	101.0	14215	Pave	Missing	IR1	Lvl	AllPub	Corner	Gtl	NridgHt	Norm
21	22	45	RM	57.0	7449	Pave	Grvl	Reg	Bnk	AllPub	Inside	Gtl	IDOTRR	Norm
22	23	20	RL	75.0	9742	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	CollgCr	Norm
23	24	120	RM	44.0	4224	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	MeadowV	Norm
24	25	20	RL	69.0	8246	Pave	Missing	IR1	Lvl	AllPub	Inside	Gtl	Sawyer	Norm
25	26	20	RL	110.0	14230	Pave	Missing	Reg	Lvl	AllPub	Corner	Gtl	NridgHt	Norm
26	27	20	RL	60.0	7200	Pave	Missing	Reg	Lvl	AllPub	Corner	Gtl	NAmes	Norm
27	28	20	RL	98.0	11478	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	NridgHt	Norm
28	29	20	RL	47.0	16321	Pave	Missing	IR1	Lvl	AllPub	CulDSac	Gtl	NAmes	Norm
29	30	30	RM	60.0	6324	Pave	Missing	IR1	Lvl	AllPub	Inside	Gtl	BrkSide	Feedr
30	31	70	C (all)	50.0	8500	Pave	Pave	Reg	Lvl	AllPub	Inside	Gtl	IDOTRR	Feedr
31	32	20	RL	69.0	8544	Pave	Missing	IR1	Lvl	AllPub	CulDSac	Gtl	Sawyer	Norm
32	33	20	RL	85.0	11049	Pave	Missing	Reg	Lvl	AllPub	Corner	Gtl	CollgCr	Norm
33	34	20	RL	70.0	10552	Pave	Missing	IR1	Lvl	AllPub	Inside	Gtl	NAmes	Norm
34	35	120	RL	60.0	7313	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	NridgHt	Norm
35	36	60	RL	108.0	13418	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	NridgHt	Norm
36	37	20	RL	112.0	10859	Pave	Missing	Reg	Lvl	AllPub	Corner	Gtl	CollgCr	Norm
37	38	20	RL	74.0	8532	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	NAmes	Norm
38	39	20	RL	68.0	7922	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	NAmes	Norm
39	40	90	RL	65.0	6040	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	Edwards	Norm
40	41	20	RL	84.0	8658	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	NAmes	Norm
41	42	20	RL	115.0	16905	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	Timber	Norm
42	43	85	RL	69.0	9180	Pave	Missing	IR1	Lvl	AllPub	CulDSac	Gtl	SawyerW	Norm
43	44	20	RL	69.0	9200	Pave	Missing	IR1	Lvl	AllPub	CulDSac	Gtl	CollgCr	Norm
44	45	20	RL	70.0	7945	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	NAmes	Norm
45	46	120	RL	61.0	7658	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	NridgHt	Norm
46	47	50	RL	48.0	12822	Pave	Missing	IR1	Lvl	AllPub	CulDSac	Gtl	Mitchel	Norm
47	48	20	FV	84.0	11096	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	Somerst	Norm

48	49	190	RM	33.0	4456	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	OldTown	Norm
49	50	20	RL	66.0	7742	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	Sawyer	Norm

In [148]:

```
## Temporal Variables (Date Time Variables)

for feature in ['YearBuilt','YearRemodAdd','GarageYrBlt']:

    dataset[feature]=dataset['YrSold']-dataset[feature]
```

In [149]:

```
dataset.head()
```

Out[149]:

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	Neighborhood	Condition1	C
0	1	60	RL	65.0	8450	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	N
1	2	20	RL	80.0	9600	Pave	Missing	Reg	Lvl	AllPub	FR2	Gtl	Veenker	Feedr	N
2	3	60	RL	68.0	11250	Pave	Missing	IR1	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	N
3	4	70	RL	60.0	9550	Pave	Missing	IR1	Lvl	AllPub	Corner	Gtl	Crawfor	Norm	N
4	5	60	RL	84.0	14260	Pave	Missing	IR1	Lvl	AllPub	FR2	Gtl	NoRidge	Norm	N

In [150]:

```
dataset[['YearBuilt','YearRemodAdd','GarageYrBlt']].head()
```

Out[150]:

	YearBuilt	YearRemodAdd	GarageYrBlt
0	5	5	5.0
1	31	31	31.0
2	7	6	7.0
3	91	36	8.0
4	8	8	8.0

Numerical Variables

Since the numerical variables are skewed we will perform log normal distribution

In [151]:

```
dataset.head()
```

Out[151]:

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	Neighborhood	Condition1	C
0	1	60	RL	65.0	8450	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	N
1	2	20	RL	80.0	9600	Pave	Missing	Reg	Lvl	AllPub	FR2	Gtl	Veenker	Feedr	N
2	3	60	RL	68.0	11250	Pave	Missing	IR1	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	N
3	4	70	RL	60.0	9550	Pave	Missing	IR1	Lvl	AllPub	Corner	Gtl	Crawfor	Norm	N
4	5	60	RL	84.0	14260	Pave	Missing	IR1	Lvl	AllPub	FR2	Gtl	NoRidge	Norm	N

```
In [152]: import numpy as np
          num_features=['LotFrontage', 'LotArea', '1stFlrSF', 'GrLivArea', 'SalePrice']

          for feature in num_features:
              dataset[feature]=np.log(dataset[feature])
```

```
In [153]: dataset.head()
```

Out[153]:

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	Neighborhood	Condition1
0	1	60	RL	4.174387	9.041922	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	CollgCr	Norm
1	2	20	RL	4.382027	9.169518	Pave	Missing	Reg	Lvl	AllPub	FR2	Gtl	Veenker	Feedr
2	3	60	RL	4.219508	9.328123	Pave	Missing	IR1	Lvl	AllPub	Inside	Gtl	CollgCr	Norm
3	4	70	RL	4.094345	9.164296	Pave	Missing	IR1	Lvl	AllPub	Corner	Gtl	Crawfor	Norm
4	5	60	RL	4.430817	9.565214	Pave	Missing	IR1	Lvl	AllPub	FR2	Gtl	NoRidge	Norm

Handling Rare Categorical Feature

We will remove categorical variables that are present less than 1% of the observations

```
In [154]: categorical_features=[feature for feature in dataset.columns if dataset[feature].dtype=='O']
```

```
In [155]: categorical_features
```

```
Out[155]: ['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',
           'Electrical',
           'KitchenQual',
           'Functional',
           'FireplaceQu',
           'GarageType',
           'GarageFinish',
           'GarageQual',
           'GarageCond',
           'PavedDrive',
           'PoolQC',
           'Fence',
           'MiscFeature',
           'SaleType',
           'SaleCondition']
```

In [156]:
```python
for feature in categorical_features:
    temp=dataset.groupby(feature)['SalePrice'].count()/len(dataset)
    temp_df=temp[temp>0.01].index
    dataset[feature]=np.where(dataset[feature].isin(temp_df),dataset[feature],'Rare_var')
```

In [157]:
```python
dataset.head(100)
```

Out[157]:

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	Neighborhood	Condition

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	Neighborhood	Condition
0	1	60	RL	4.174387	9.041922	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	CollgCr	Norm
1	2	20	RL	4.382027	9.169518	Pave	Missing	Reg	Lvl	AllPub	FR2	Gtl	Rare_var	Feedr
2	3	60	RL	4.219508	9.328123	Pave	Missing	IR1	Lvl	AllPub	Inside	Gtl	CollgCr	Norm
3	4	70	RL	4.094345	9.164296	Pave	Missing	IR1	Lvl	AllPub	Corner	Gtl	Crawfor	Norm
4	5	60	RL	4.430817	9.565214	Pave	Missing	IR1	Lvl	AllPub	FR2	Gtl	NoRidge	Norm
5	6	50	RL	4.442651	9.554993	Pave	Missing	IR1	Lvl	AllPub	Inside	Gtl	Mitchel	Norm
6	7	20	RL	4.317488	9.218705	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	Somerst	Norm
7	8	60	RL	4.234107	9.247829	Pave	Missing	IR1	Lvl	AllPub	Corner	Gtl	NWAmes	PosN
8	9	50	RM	3.931826	8.719317	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	OldTown	Artery
9	10	190	RL	3.912023	8.911934	Pave	Missing	Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Artery
10	11	20	RL	4.248495	9.323669	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	Sawyer	Norm
11	12	60	RL	4.442651	9.386308	Pave	Missing	IR1	Lvl	AllPub	Inside	Gtl	NridgHt	Norm
12	13	20	RL	4.234107	9.470240	Pave	Missing	IR2	Lvl	AllPub	Inside	Gtl	Sawyer	Norm
13	14	20	RL	4.510860	9.273503	Pave	Missing	IR1	Lvl	AllPub	Inside	Gtl	CollgCr	Norm
14	15	20	RL	4.234107	9.298351	Pave	Missing	IR1	Lvl	AllPub	Corner	Gtl	NAmes	Norm
15	16	45	RM	3.931826	8.719317	Pave	Missing	Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Norm
16	17	20	RL	4.234107	9.327323	Pave	Missing	IR1	Lvl	AllPub	CulDSac	Gtl	NAmes	Norm
17	18	90	RL	4.276666	9.286468	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	Sawyer	Norm
18	19	20	RL	4.189655	9.524786	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	SawyerW	Rare_var
19	20	20	RL	4.248495	8.930626	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	NAmes	Norm
20	21	60	RL	4.615121	9.562053	Pave	Missing	IR1	Lvl	AllPub	Corner	Gtl	NridgHt	Norm
21	22	45	RM	4.043051	8.915835	Pave	Grvl	Reg	Bnk	AllPub	Inside	Gtl	IDOTRR	Norm
22	23	20	RL	4.317488	9.184202	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	CollgCr	Norm
23	24	120	RM	3.784190	8.348538	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	MeadowV	Norm
24	25	20	RL	4.234107	9.017484	Pave	Missing	IR1	Lvl	AllPub	Inside	Gtl	Sawyer	Norm
25	26	20	RL	4.700480	9.563108	Pave	Missing	Reg	Lvl	AllPub	Corner	Gtl	NridgHt	Norm
26	27	20	RL	4.094345	8.881836	Pave	Missing	Reg	Lvl	AllPub	Corner	Gtl	NAmes	Norm
27	28	20	RL	4.584967	9.348187	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	NridgHt	Norm

27	28	20	RL	4.584967	9.348187	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	NridgHt	Norm
28	29	20	RL	3.850148	9.700208	Pave	Missing	IR1	Lvl	AllPub	CulDSac	Gtl	NAmes	Norm
29	30	30	RM	4.094345	8.752107	Pave	Missing	IR1	Lvl	AllPub	Inside	Gtl	BrkSide	Feedr
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
70	71	20	RL	4.553877	9.521568	Pave	Missing	IR1	Lvl	AllPub	Inside	Gtl	NAmes	Norm
71	72	20	RL	4.234107	8.935772	Pave	Missing	Reg	Lvl	AllPub	Corner	Gtl	Mitchel	Norm
72	73	60	RL	4.304065	9.224342	Pave	Missing	IR1	Lvl	AllPub	Corner	Gtl	Gilbert	Norm
73	74	20	RL	4.442651	9.230143	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	NAmes	Norm
74	75	50	RM	4.094345	8.663888	Pave	Missing	Reg	Lvl	AllPub	Corner	Gtl	OldTown	Norm
75	76	180	RM	3.044522	7.375256	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	MeadowV	Norm
76	77	20	RL	4.234107	9.044876	Pave	Missing	IR1	Lvl	AllPub	Inside	Gtl	NAmes	Norm
77	78	50	RM	3.912023	9.063579	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	BrkSide	Norm
78	79	90	RL	4.276666	9.285262	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	Sawyer	Norm
79	80	50	RM	4.094345	9.253400	Pave	Grvl	Reg	Lvl	AllPub	Corner	Gtl	OldTown	Norm
80	81	60	RL	4.605170	9.472705	Pave	Missing	Reg	Lvl	AllPub	Corner	Gtl	NAmes	Norm
81	82	120	RM	3.465736	8.411833	Pave	Missing	Reg	Lvl	AllPub	FR2	Gtl	Mitchel	Norm
82	83	20	RL	4.356709	9.230731	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	Somerst	Norm
83	84	20	RL	4.382027	9.092907	Pave	Missing	IR1	Lvl	AllPub	Inside	Gtl	NAmes	Norm
84	85	80	RL	4.234107	9.051345	Pave	Missing	IR1	Lvl	AllPub	Inside	Gtl	Gilbert	Norm
85	86	60	RL	4.795791	9.684025	Pave	Missing	Reg	Lvl	AllPub	Corner	Gtl	NoRidge	Norm
86	87	60	RL	4.804021	9.385218	Pave	Missing	IR2	Lvl	AllPub	Inside	Gtl	Gilbert	Norm
87	88	160	FV	3.688879	8.281724	Pave	Pave	Reg	Lvl	AllPub	Corner	Gtl	Somerst	Norm
88	89	50	Rare_var	4.653960	9.044286	Pave	Missing	IR1	Lvl	AllPub	Corner	Gtl	IDOTRR	Feedr
89	90	20	RL	4.094345	8.995909	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	CollgCr	Norm
90	91	20	RL	4.094345	8.881836	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	NAmes	Norm
91	92	20	RL	4.442651	9.047821	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	NAmes	Norm
92	93	30	RL	4.382027	9.500020	Pave	Grvl	IR1	HLS	AllPub	Inside	Gtl	Crawfor	Norm
93	94	190	Rare_var	4.094345	8.881836	Pave	Missing	Reg	Lvl	AllPub	Corner	Gtl	OldTown	Norm
94	95	60	RL	4.234107	9.141740	Pave	Missing	IR1	Lvl	AllPub	Inside	Gtl	CollgCr	Norm

94	95	60	RL	4.234107	9.141740	Pave	Missing	IR1	Lvl	AllPub	Inside	Gtl	CollgCr	Norm
95	96	60	RL	4.234107	9.186560	Pave	Missing	IR2	Lvl	AllPub	Corner	Gtl	Gilbert	Norm
96	97	20	RL	4.356709	9.236398	Pave	Missing	IR1	Lvl	AllPub	Inside	Gtl	CollgCr	Norm
97	98	20	RL	4.290459	9.298443	Pave	Missing	Reg	HLS	AllPub	Inside	Gtl	Edwards	Norm
98	99	30	RL	4.442651	9.270965	Pave	Missing	Reg	Lvl	AllPub	Corner	Gtl	Edwards	Norm
99	100	20	RL	4.343805	9.139918	Pave	Missing	IR1	Lvl	AllPub	Inside	Gtl	NAmes	Norm

100 rows × 84 columns

In [ ]:

In [158]:

```
for feature in categorical_features:
    labels_ordered=dataset.groupby([feature])['SalePrice'].mean().sort_values().index
    labels_ordered={k:i for i,k in enumerate(labels_ordered,0)}
    dataset[feature]=dataset[feature].map(labels_ordered)
```

In [159]:

```
dataset.head(10)
```

Out[159]:

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	Neighborhood	Condition1	Co
0	1	60	3	4.174387	9.041922	1	2	0	1	1	0	0	14	2	1
1	2	20	3	4.382027	9.169518	1	2	0	1	1	2	0	11	1	1
2	3	60	3	4.219508	9.328123	1	2	1	1	1	0	0	14	2	1
3	4	70	3	4.094345	9.164296	1	2	1	1	1	1	0	16	2	1
4	5	60	3	4.430817	9.565214	1	2	1	1	1	2	0	22	2	1
5	6	50	3	4.442651	9.554993	1	2	1	1	1	0	0	9	2	1
6	7	20	3	4.317488	9.218705	1	2	0	1	1	0	0	18	2	1
7	8	60	3	4.234107	9.247829	1	2	1	1	1	1	0	12	5	1
8	9	50	1	3.931826	8.719317	1	2	0	1	1	0	0	4	0	1
9	10	190	3	3.912023	8.911934	1	2	0	1	1	1	0	3	0	0

In [160]:

```
scaling_feature=[feature for feature in dataset.columns if feature not in ['Id','SalePerice'] ]
len(scaling_feature)
```

Out[160]: 83

```
In [161]: scaling_feature
```

```
Out[161]: ['MSSubClass',
           'MSZoning',
           'LotFrontage',
           'LotArea',
           'Street',
           'Alley',
           'LotShape',
           'LandContour',
           'Utilities',
           'LotConfig',
           'LandSlope',
           'Neighborhood',
           'Condition1',
           'Condition2',
           'BldgType',
           'HouseStyle',
           'OverallQual',
           'OverallCond',
           'YearBuilt',
           'YearRemodAdd',
           'RoofStyle',
           'RoofMatl',
           'Exterior1st',
           'Exterior2nd',
           'MasVnrType',
           'MasVnrArea',
           'ExterQual',
           'ExterCond',
           'Foundation',
           'BsmtQual',
           'BsmtCond',
           'BsmtExposure',
           'BsmtFinType1',
           'BsmtFinSF1',
           'BsmtFinType2',
           'BsmtFinSF2',
           'BsmtUnfSF',
           'TotalBsmtSF',
           'Heating',
           'HeatingQC',
           'CentralAir',
           'Electrical',
           '1stFlrSF',
           '2ndFlrSF',
           'LowQualFinSF',
 'LowQualFinSF',
 'GrLivArea',
 'BsmtFullBath',
 'BsmtHalfBath',
 'FullBath',
 'HalfBath',
 'BedroomAbvGr',
 'KitchenAbvGr',
 'KitchenQual',
 'TotRmsAbvGrd',
 'Functional',
 'Fireplaces',
 'FireplaceQu',
 'GarageType',
 'GarageYrBlt',
 'GarageFinish',
 'GarageCars',
 'GarageArea',
 'GarageQual',
 'GarageCond',
 'PavedDrive',
 'WoodDeckSF',
 'OpenPorchSF',
 'EnclosedPorch',
 '3SsnPorch',
 'ScreenPorch',
 'PoolArea',
 'PoolQC',
 'Fence',
 'MiscFeature',
 'MiscVal',
 'MoSold',
 'YrSold',
 'SaleType',
 'SaleCondition',
 'SalePrice',
 'LotFrontagenan',
 'MasVnrAreanan',
 'GarageYrBltnan']
```

In [162]:
```
dataset.head()
```

Out[162]:

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	Neighborhood	Condition1	Cc
0	1	60	3	4.174387	9.041922	1	2	0	1	1	0	0	14	2	1
1	2	20	3	4.382027	9.169518	1	2	0	1	1	2	0	11	1	1
2	3	60	3	4.219508	9.328123	1	2	1	1	1	0	0	14	2	1

3	4	70	3	4.094345	9.164296	1	2	1	1	1	1	0	16	2	1
4	5	60	3	4.430817	9.565214	1	2	1	1	1	2	0	22	2	1

Feature Scaling

In [166]:

```
feature_scale=[feature for feature in dataset.columns if feature not in ['Id','SalePrice']]

from sklearn.preprocessing import MinMaxScaler
scaler=MinMaxScaler()
scaler.fit(dataset[feature_scale])
```

```
C:\Users\krish.naik\AppData\Local\Continuum\anaconda3\lib\site-packages\sklearn\preprocessing\data.py:323: DataConversionWarning: Data w
ith input dtype int32, int64, float64 were all converted to float64 by MinMaxScaler.
  return self.partial_fit(X, y)
```

Out[166]:

```
MinMaxScaler(copy=True, feature_range=(0, 1))
```

In [168]:

```
scaler.transform(dataset[feature_scale])
```

Out[168]:

```
array([[0.23529412, 0.75      , 0.41820812, ..., 0.        , 0.        ,
        0.        ],
       [0.        , 0.75      , 0.49506375, ..., 0.        , 0.        ,
        0.        ],
       [0.23529412, 0.75      , 0.434909  , ..., 0.        , 0.        ,
        0.        ],
       ...,
       [0.29411765, 0.75      , 0.42385922, ..., 0.        , 0.        ,
        0.        ],
       [0.        , 0.75      , 0.434909  , ..., 0.        , 0.        ,
        0.        ],
       [0.        , 0.75      , 0.47117546, ..., 0.        , 0.        ,
        0.        ]])
```

In [169]:

```
# transform the train and test set, and add on the Id and SalePrice variables
data = pd.concat([dataset[['Id', 'SalePrice']].reset_index(drop=True),
                    pd.DataFrame(scaler.transform(dataset[feature_scale]), columns=feature_scale)],
                    axis=1)
```

In [170]:

```
data.head()
```

Out[170]:

	Id	SalePrice	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	Neighborhood	Cor
0	1	12.247694	0.235294	0.75	0.418208	0.366344	1.0	1.0	0.000000	0.333333	1.0	0.00	0.0	0.636364	0.4

1	2	12.109011	0.000000	0.75	0.495064	0.391317	1.0	1.0	0.000000	0.333333	1.0	0.50	0.0	0.500000	0.2
2	3	12.317167	0.235294	0.75	0.434909	0.422359	1.0	1.0	0.333333	0.333333	1.0	0.00	0.0	0.636364	0.4
3	4	11.849398	0.294118	0.75	0.388581	0.390295	1.0	1.0	0.333333	0.333333	1.0	0.25	0.0	0.727273	0.4
4	5	12.429216	0.235294	0.75	0.513123	0.468761	1.0	1.0	0.333333	0.333333	1.0	0.50	0.0	1.000000	0.4