Data Preprocessing

Loading dataset.

In [1]:
```
import numpy as np
import pandas as pd
```

In [2]:
```
df = pd.read_csv('Data.csv')

df
```

Out[2]:

	Country	Age	Salary	Purchased
0	France	44.0	72000.0	No
1	Spain	27.0	48000.0	Yes
2	Germany	30.0	54000.0	No
3	Spain	38.0	61000.0	No
4	Germany	40.0	NaN	Yes
5	France	35.0	58000.0	Yes
6	Spain	NaN	52000.0	No
7	France	48.0	79000.0	Yes
8	Germany	50.0	83000.0	No
9	France	37.0	67000.0	Yes

In [3]:
```
df.shape
```

Out[3]:
```
(10, 4)
```

In [4]:
```
df.dtypes
```

Out[4]:
```
Country       object
Age          float64
Salary       float64
Purchased     object
dtype: object
```

Observations :

- There are 2 categorical data type and 2 numerical data type.

In [5]:
```
df.isnull().sum()
```

Out[5]:
```
Country      0
Age          1
Salary       1
Purchased    0
dtype: int64
```

Observations :

- There are some null values in both the numerical features. SO, we can handle them by replacing them with mean.

In [6]:
```
df.head()
```

Out[6]:

	Country	Age	Salary	Purchased
0	France	44.0	72000.0	No
1	Spain	27.0	48000.0	Yes
2	Germany	30.0	54000.0	No
3	Spain	38.0	61000.0	No
4	Germany	40.0	NaN	Yes

Taking care of missing values.

In [7]:
```
# handling independent numerical features
from sklearn.impute import SimpleImputer

numerical_features = [feature for feature in df.columns if df[feature].dtype != 'O']

imputer = SimpleImputer(missing_values = np.nan, strategy = 'mean')
df[numerical_features] = imputer.fit_transform(df[numerical_features])

df
```

Out[7]:

	Country	Age	Salary	Purchased
0	France	44.000000	72000.000000	No
1	Spain	27.000000	48000.000000	Yes
2	Germany	30.000000	54000.000000	No
3	Spain	38.000000	61000.000000	No
4	Germany	40.000000	63777.777778	Yes
5	France	35.000000	58000.000000	Yes
6	Spain	38.777778	52000.000000	No
7	France	48.000000	79000.000000	Yes
8	Germany	50.000000	83000.000000	No
9	France	37.000000	67000.000000	Yes

In [18]:
```
df.isnull().sum()
```

Out[18]:
```
Age          0
Salary       0
Purchased    0
dtype: int64
```

Observations :

- There is no missing values.

Taking care of categorical features.

In [8]:
```
# handling independent categorical features
from sklearn.preprocessing import OneHotEncoder

encoder = OneHotEncoder()
encoded = encoder.fit_transform(df[['Country']]).toarray()
encoded_df = pd.DataFrame(encoded, columns = encoder.get_feature_names_out())

df.drop(columns = ['Country'], inplace = True)
new_df = pd.concat([df, encoded_df], axis = 1)

new_df
```

Out[8]:

	Age	Salary	Purchased	Country_France	Country_Germany	Country_Spain
0	44.000000	72000.000000	No	1.0	0.0	0.0
1	27.000000	48000.000000	Yes	0.0	0.0	1.0
2	30.000000	54000.000000	No	0.0	1.0	0.0
3	38.000000	61000.000000	No	0.0	0.0	1.0
4	40.000000	63777.777778	Yes	0.0	1.0	0.0
5	35.000000	58000.000000	Yes	1.0	0.0	0.0
6	38.777778	52000.000000	No	0.0	0.0	1.0
7	48.000000	79000.000000	Yes	1.0	0.0	0.0
8	50.000000	83000.000000	No	0.0	1.0	0.0
9	37.000000	67000.000000	Yes	1.0	0.0	0.0

In [9]:
```
# handling dependent categorical features
from sklearn.preprocessing import LabelEncoder

encoder = LabelEncoder()
new_df['Purchased'] = encoder.fit_transform(new_df['Purchased'])

new_df
```

Out[9]:

	Age	Salary	Purchased	Country_France	Country_Germany	Country_Spain
0	44.000000	72000.000000	0	1.0	0.0	0.0
1	27.000000	48000.000000	1	0.0	0.0	1.0
2	30.000000	54000.000000	0	0.0	1.0	0.0
3	38.000000	61000.000000	0	0.0	0.0	1.0
4	40.000000	63777.777778	1	0.0	1.0	0.0
5	35.000000	58000.000000	1	1.0	0.0	0.0
6	38.777778	52000.000000	0	0.0	0.0	1.0
7	48.000000	79000.000000	1	1.0	0.0	0.0
8	50.000000	83000.000000	0	0.0	1.0	0.0
9	37.000000	67000.000000	1	1.0	0.0	0.0

Splitting dataset into independent and dependent features

In [10]:
```
from sklearn.model_selection import train_test_split

x = new_df.drop(columns = ['Purchased'])
y = new_df['Purchased']

x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.2, random_state = 42)
```

In [11]:
```
x_train
```

Out[11]:

	Age	Salary	Country_France	Country_Germany	Country_Spain
5	35.000000	58000.000000	1.0	0.0	0.0
0	44.000000	72000.000000	1.0	0.0	0.0
7	48.000000	79000.000000	1.0	0.0	0.0
2	30.000000	54000.000000	0.0	1.0	0.0
9	37.000000	67000.000000	1.0	0.0	0.0
4	40.000000	63777.777778	0.0	1.0	0.0
3	38.000000	61000.000000	0.0	0.0	1.0
6	38.777778	52000.000000	0.0	0.0	1.0

In [12]:
```
y_train
```

Out[12]:
```
5    1
0    0
7    1
2    0
9    1
4    1
3    0
6    0
Name: Purchased, dtype: int32
```

In [13]:
```
x_test
```

Out[13]:

	Age	Salary	Country_France	Country_Germany	Country_Spain
8	50.0	83000.0	0.0	1.0	0.0
1	27.0	48000.0	0.0	0.0	1.0

In [14]:
```
y_test
```

Out[14]:
```
8    0
1    1
Name: Purchased, dtype: int32
```

Standardization

In [15]:
```
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()
x_train = scaler.fit_transform(x_train)
x_test = scaler.transform(x_test) # to avoid data leakage we just do transform
```

In [16]:
```
x_train
```

Out[16]:
```
array([[-0.7529426 , -0.62603778,  1.        , -0.57735027, -0.57735027],
       [ 1.00845381,  1.01304295,  1.        , -0.57735027, -0.57735027],
       [ 1.79129666,  1.83258331,  1.        , -0.57735027, -0.57735027],
       [-1.73149616, -1.09434656, -1.        ,  1.73205081, -0.57735027],
       [-0.36152118,  0.42765698,  1.        , -0.57735027, -0.57735027],
       [ 0.22561096,  0.05040824, -1.        ,  1.73205081, -0.57735027],
       [-0.16581046, -0.27480619, -1.        , -0.57735027,  1.73205081],
       [-0.01359102, -1.32850095, -1.        , -0.57735027,  1.73205081]])
```

In [17]:
```
x_test
```

Out[17]:
```
array([[ 2.18271808,  2.30089209, -1.        ,  1.73205081, -0.57735027],
       [-2.3186283 , -1.79680973, -1.        , -0.57735027,  1.73205081]])
```

Advance House Price Prediction - Feature Engineering

---

we will performing the all the below steps in feature engineering

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

In [19]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns

pd.pandas.set_option('display.max_columns', None)
pd.pandas.set_option('display.max_rows', None)
```

In [2]:
```
dataset = pd.read_csv('Advance_House_Price_Prediction_train.csv')
dataset.head()
```

Out[2]:

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	Neighborhood	Condition1	Condition2	BldgType	HouseStyle	OverallQual	OverallCond	YearBuilt	YearRemodAdd	RoofStyle	RoofMatl	Exterior1st	E
0	1	60	RL	65.0	8450	Pave	NaN	Reg	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	2Story	7	5	2003	2003	Gable	CompShg	VinylSd	
1	2	20	RL	80.0	9600	Pave	NaN	Reg	Lvl	AllPub	FR2	Gtl	Veenker	Feedr	Norm	1Fam	1Story	6	8	1976	1976	Gable	CompShg	MetalSd	
2	3	60	RL	68.0	11250	Pave	NaN	IR1	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	2Story	7	5	2001	2002	Gable	CompShg	VinylSd	
3	4	70	RL	60.0	9550	Pave	NaN	IR1	Lvl	AllPub	Corner	Gtl	Crawfor	Norm	Norm	1Fam	2Story	7	5	1915	1970	Gable	CompShg	Wd Sdng	
4	5	60	RL	84.0	14260	Pave	NaN	IR1	Lvl	AllPub	FR2	Gtl	NoRidge	Norm	Norm	1Fam	2Story	8	5	2000	2000	Gable	CompShg	VinylSd	

In [3]:
```
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 = 42)
```

In [4]:
```
x_train.shape, x_test.shape
```

Out[4]:
```
((1314, 81), (146, 81))
```

Missing Values

In [5]:
```
# 1. categorical feature
feature_nan = [feature for feature in dataset.columns if (dataset[feature].isnull().sum() > 1 and dataset[feature].dtypes == 'O')]

for feature in feature_nan:
    print(f'{feature} categorical feature has {np.round(dataset[feature].isnull().mean() * 100, 2)}%')
```

```
Alley categorical feature has 93.77%
MasVnrType categorical feature has 0.55%
BsmtQual categorical feature has 2.53%
BsmtCond categorical feature has 2.53%
BsmtExposure categorical feature has 2.6%
BsmtFinType1 categorical feature has 2.53%
BsmtFinType2 categorical feature has 2.6%
FireplaceQu categorical feature has 47.26%
GarageType categorical feature has 5.55%
GarageFinish categorical feature has 5.55%
GarageQual categorical feature has 5.55%
GarageCond categorical feature has 5.55%
PoolQC categorical feature has 99.52%
Fence categorical feature has 80.75%
MiscFeature categorical feature has 96.3%
```

In [6]:
```
# replacing missing values 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, feature_nan)

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

Out[6]:
```
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 [7]:
```
dataset.head()
```

Out[7]:

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	Neighborhood	Condition1	Condition2	BldgType	HouseStyle	OverallQual	OverallCond	YearBuilt	YearRemodAdd	RoofStyle	RoofMatl	Exterior1st
0	1	60	RL	65.0	8450	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	2Story	7	5	2003	2003	Gable	CompShg	VinylSd
1	2	20	RL	80.0	9600	Pave	Missing	Reg	Lvl	AllPub	FR2	Gtl	Veenker	Feedr	Norm	1Fam	1Story	6	8	1976	1976	Gable	CompShg	MetalSd
2	3	60	RL	68.0	11250	Pave	Missing	IR1	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	2Story	7	5	2001	2002	Gable	CompShg	VinylSd
3	4	70	RL	60.0	9550	Pave	Missing	IR1	Lvl	AllPub	Corner	Gtl	Crawfor	Norm	Norm	1Fam	2Story	7	5	1915	1970	Gable	CompShg	Wd Sdng
4	5	60	RL	84.0	14260	Pave	Missing	IR1	Lvl	AllPub	FR2	Gtl	NoRidge	Norm	Norm	1Fam	2Story	8	5	2000	2000	Gable	CompShg	VinylSd

In [8]:
```
# 2. numerical value
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.round(dataset[feature].isnull().mean() * 100,2)))
```

```
LotFrontage: 17.74% missing value
MasVnrArea: 0.55% missing value
GarageYrBlt: 5.55% missing value
```

In [9]:
```
# replacing nan numberical feature with the median values
for feature in numerical_with_nan:
    median_value = dataset[feature].median()
    dataset[feature + 'nan'] = np.where(dataset[feature].isnull(), 1, 0)
    dataset[feature].fillna(median_value, inplace = True)

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

Out[9]:
```
LotFrontage    0
MasVnrArea     0
GarageYrBlt    0
dtype: int64
```

In [10]:
```
dataset.head(20)
```

Out[10]:

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	Neighborhood	Condition1	Condition2	BldgType	HouseStyle	OverallQual	OverallCond	YearBuilt	YearRemodAdd	RoofStyle	RoofMatl	Exterior1s
0	1	60	RL	65.0	8450	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	2Story	7	5	2003	2003	Gable	CompShg	VinylSc
1	2	20	RL	80.0	9600	Pave	Missing	Reg	Lvl	AllPub	FR2	Gtl	Veenker	Feedr	Norm	1Fam	1Story	6	8	1976	1976	Gable	CompShg	MetalSc
2	3	60	RL	68.0	11250	Pave	Missing	IR1	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	2Story	7	5	2001	2002	Gable	CompShg	VinylSc
3	4	70	RL	60.0	9550	Pave	Missing	IR1	Lvl	AllPub	Corner	Gtl	Crawfor	Norm	Norm	1Fam	2Story	7	5	1915	1970	Gable	CompShg	Wd Sdng
4	5	60	RL	84.0	14260	Pave	Missing	IR1	Lvl	AllPub	FR2	Gtl	NoRidge	Norm	Norm	1Fam	2Story	8	5	2000	2000	Gable	CompShg	VinylSc
5	6	50	RL	85.0	14115	Pave	Missing	IR1	Lvl	AllPub	Inside	Gtl	Mitchel	Norm	Norm	1Fam	1.5Fin	5	5	1993	1995	Gable	CompShg	VinylSc
6	7	20	RL	75.0	10084	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	Somerst	Norm	Norm	1Fam	1Story	8	5	2004	2005	Gable	CompShg	VinylSc
7	8	60	RL	69.0	10382	Pave	Missing	IR1	Lvl	AllPub	Corner	Gtl	NWAmes	PosN	Norm	1Fam	2Story	7	6	1973	1973	Gable	CompShg	HdBoarc
8	9	50	RM	51.0	6120	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	OldTown	Artery	Norm	1Fam	1.5Fin	7	5	1931	1950	Gable	CompShg	BrkFace
9	10	190	RL	50.0	7420	Pave	Missing	Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Artery	Artery	2fmCon	1.5Unf	5	6	1939	1950	Gable	CompShg	MetalSc
10	11	20	RL	70.0	11200	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	Sawyer	Norm	Norm	1Fam	1Story	5	5	1965	1965	Hip	CompShg	HdBoarc
11	12	60	RL	85.0	11924	Pave	Missing	IR1	Lvl	AllPub	Inside	Gtl	NridgHt	Norm	Norm	1Fam	2Story	9	5	2005	2006	Hip	CompShg	WdShing
12	13	20	RL	69.0	12968	Pave	Missing	IR2	Lvl	AllPub	Inside	Gtl	Sawyer	Norm	Norm	1Fam	1Story	5	6	1962	1962	Hip	CompShg	HdBoarc
13	14	20	RL	91.0	10652	Pave	Missing	IR1	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	1Story	7	5	2006	2007	Gable	CompShg	VinylSc
14	15	20	RL	69.0	10920	Pave	Missing	IR1	Lvl	AllPub	Corner	Gtl	NAmes	Norm	Norm	1Fam	1Story	6	5	1960	1960	Hip	CompShg	MetalSc
15	16	45	RM	51.0	6120	Pave	Missing	Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Norm	Norm	1Fam	1.5Unf	7	8	1929	2001	Gable	CompShg	Wd Sdng
16	17	20	RL	69.0	11241	Pave	Missing	IR1	Lvl	AllPub	CulDSac	Gtl	NAmes	Norm	Norm	1Fam	1Story	6	7	1970	1970	Gable	CompShg	Wd Sdng
17	18	90	RL	72.0	10791	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	Sawyer	Norm	Norm	Duplex	1Story	4	5	1967	1967	Gable	CompShg	MetalSc
18	19	20	RL	66.0	13695	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	SawyerW	RRAe	Norm	1Fam	1Story	5	5	2004	2004	Gable	CompShg	VinylSc
19	20	20	RL	70.0	7560	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	NAmes	Norm	Norm	1Fam	1Story	5	6	1958	1965	Hip	CompShg	BrkFace

In [11]:
```
# 3. temperal variables - datetime variables

# from eda we have observed that year sold with sale price was decreasing
year_feature = [feature for feature in dataset.columns if ('Yr' in feature or 'Year' in feature)]

year_feature
```

Out[11]:
```
['YearBuilt', 'YearRemodAdd', 'GarageYrBlt', 'YrSold', 'GarageYrBltnan']
```

In [12]:
```
for feature in year_feature:
    if ('YrSold' not in feature) and ('GarageYrBltnan' not in feature):
        dataset[feature] = dataset['YrSold'] - dataset[feature]

dataset.head()
```

Out[12]:

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	Neighborhood	Condition1	Condition2	BldgType	HouseStyle	OverallQual	OverallCond	YearBuilt	YearRemodAdd	RoofStyle	RoofMatl	Exterior1st
0	1	60	RL	65.0	8450	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	2Story	7	5	5	5	Gable	CompShg	VinylSd
1	2	20	RL	80.0	9600	Pave	Missing	Reg	Lvl	AllPub	FR2	Gtl	Veenker	Feedr	Norm	1Fam	1Story	6	8	31	31	Gable	CompShg	MetalSd
2	3	60	RL	68.0	11250	Pave	Missing	IR1	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	2Story	7	5	7	6	Gable	CompShg	VinylSd
3	4	70	RL	60.0	9550	Pave	Missing	IR1	Lvl	AllPub	Corner	Gtl	Crawfor	Norm	Norm	1Fam	2Story	7	5	91	36	Gable	CompShg	Wd Sdng
4	5	60	RL	84.0	14260	Pave	Missing	IR1	Lvl	AllPub	FR2	Gtl	NoRidge	Norm	Norm	1Fam	2Story	8	5	8	8	Gable	CompShg	VinylSd

Numerical Variables

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

In [13]:
```
dataset.head()
```

Out[13]:

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	Neighborhood	Condition1	Condition2	BldgType	HouseStyle	OverallQual	OverallCond	YearBuilt	YearRemodAdd	RoofStyle	RoofMatl	Exterior1st
0	1	60	RL	65.0	8450	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	2Story	7	5	5	5	Gable	CompShg	VinylSd
1	2	20	RL	80.0	9600	Pave	Missing	Reg	Lvl	AllPub	FR2	Gtl	Veenker	Feedr	Norm	1Fam	1Story	6	8	31	31	Gable	CompShg	MetalSd
2	3	60	RL	68.0	11250	Pave	Missing	IR1	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	2Story	7	5	7	6	Gable	CompShg	VinylSd
3	4	70	RL	60.0	9550	Pave	Missing	IR1	Lvl	AllPub	Corner	Gtl	Crawfor	Norm	Norm	1Fam	2Story	7	5	91	36	Gable	CompShg	Wd Sdng
4	5	60	RL	84.0	14260	Pave	Missing	IR1	Lvl	AllPub	FR2	Gtl	NoRidge	Norm	Norm	1Fam	2Story	8	5	8	8	Gable	CompShg	VinylSd

In [14]:
```
import numpy as np
num_features = ['LotFrontage', 'LotArea', '1stFlrSF', 'GrLivArea', 'SalePrice']

# converting numerical features into log normal distribution
for feature in num_features:
    dataset[feature]=np.log(dataset[feature])
```

In [15]:
```
dataset.head()
```

Out[15]:

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	Neighborhood	Condition1	Condition2	BldgType	HouseStyle	OverallQual	OverallCond	YearBuilt	YearRemodAdd	RoofStyle	RoofMatl	Exterior1st
0	1	60	RL	4.174387	9.041922	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	2Story	7	5	5	5	Gable	CompShg	VinylSd
1	2	20	RL	4.382027	9.169518	Pave	Missing	Reg	Lvl	AllPub	FR2	Gtl	Veenker	Feedr	Norm	1Fam	1Story	6	8	31	31	Gable	CompShg	MetalSd
2	3	60	RL	4.219508	9.328123	Pave	Missing	IR1	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	2Story	7	5	7	6	Gable	CompShg	VinylSd
3	4	70	RL	4.094345	9.164296	Pave	Missing	IR1	Lvl	AllPub	Corner	Gtl	Crawfor	Norm	Norm	1Fam	2Story	7	5	91	36	Gable	CompShg	Wd Sdng
4	5	60	RL	4.430817	9.565214	Pave	Missing	IR1	Lvl	AllPub	FR2	Gtl	NoRidge	Norm	Norm	1Fam	2Story	8	5	8	8	Gable	CompShg	VinylSd

Handling Rare Categorical feature

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

In [16]:
```
categorical_features = [feature for feature in dataset.columns if dataset[feature].dtype == 'O']

categorical_features
```

Out[16]:
```
['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 [17]:
```
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 [20]:
```
dataset.head(100)
```

Out[20]:

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	Neighborhood	Condition1	Condition2	BldgType	HouseStyle	OverallQual	OverallCond	YearBuilt	YearRemodAdd	RoofStyle	RoofMatl	Exterio
0	1	60	RL	4.174387	9.041922	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	2Story	7	5	5	5	Gable	CompShg	Viny
1	2	20	RL	4.382027	9.169518	Pave	Missing	Reg	Lvl	AllPub	FR2	Gtl	Rare Var	Feedr	Norm	1Fam	1Story	6	8	31	31	Gable	CompShg	Meta
2	3	60	RL	4.219508	9.328123	Pave	Missing	IR1	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	2Story	7	5	7	6	Gable	CompShg	Viny
3	4	70	RL	4.094345	9.164296	Pave	Missing	IR1	Lvl	AllPub	Corner	Gtl	Crawfor	Norm	Norm	1Fam	2Story	7	5	91	36	Gable	CompShg	Wd S
4	5	60	RL	4.430817	9.565214	Pave	Missing	IR1	Lvl	AllPub	FR2	Gtl	NoRidge	Norm	Norm	1Fam	2Story	8	5	8	8	Gable	CompShg	Viny
5	6	50	RL	4.442651	9.554993	Pave	Missing	IR1	Lvl	AllPub	Inside	Gtl	Mitchel	Norm	Norm	1Fam	1.5Fin	5	5	16	14	Gable	CompShg	Viny
6	7	20	RL	4.317488	9.218705	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	Somerst	Norm	Norm	1Fam	1Story	8	5	3	2	Gable	CompShg	Viny
7	8	60	RL	4.234107	9.247829	Pave	Missing	IR1	Lvl	AllPub	Corner	Gtl	NWAmes	PosN	Norm	1Fam	2Story	7	6	36	36	Gable	CompShg	HdBo
8	9	50	RM	3.931826	8.719317	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	OldTown	Artery	Norm	1Fam	1.5Fin	7	5	77	58	Gable	CompShg	BrkF
9	10	190	RL	3.912023	8.911934	Pave	Missing	Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Artery	Rare Var	2fmCon	Rare Var	5	6	69	58	Gable	CompShg	Meta
10	11	20	RL	4.248495	9.323669	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	Sawyer	Norm	Norm	1Fam	1Story	5	5	43	43	Hip	CompShg	HdBo
11	12	60	RL	4.442651	9.386308	Pave	Missing	IR1	Lvl	AllPub	Inside	Gtl	NridgHt	Norm	Norm	1Fam	2Story	9	5	1	0	Hip	CompShg	WdSh
12	13	20	RL	4.234107	9.470240	Pave	Missing	IR2	Lvl	AllPub	Inside	Gtl	Sawyer	Norm	Norm	1Fam	1Story	5	6	46	46	Hip	CompShg	HdBo
13	14	20	RL	4.510860	9.273503	Pave	Missing	IR1	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	1Story	7	5	1	0	Gable	CompShg	Viny
14	15	20	RL	4.234107	9.298351	Pave	Missing	IR1	Lvl	AllPub	Corner	Gtl	NAmes	Norm	Norm	1Fam	1Story	6	5	48	48	Hip	CompShg	Meta
15	16	45	RM	3.931826	8.719317	Pave	Missing	Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Norm	Norm	1Fam	Rare Var	7	8	78	6	Gable	CompShg	Wd S
16	17	20	RL	4.234107	9.327323	Pave	Missing	IR1	Lvl	AllPub	CulDSac	Gtl	NAmes	Norm	Norm	1Fam	1Story	6	7	40	40	Gable	CompShg	Wd S
17	18	90	RL	4.276666	9.286468	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	Sawyer	Norm	Norm	Duplex	1Story	4	5	39	39	Gable	CompShg	Meta
18	19	20	RL	4.189655	9.524786	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	SawyerW	Rare Var	Norm	1Fam	1Story	5	5	4	4	Gable	CompShg	Viny
19	20	20	RL	4.248495	8.930626	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	NAmes	Norm	Norm	1Fam	1Story	5	6	51	44	Hip	CompShg	BrkF
20	21	60	RL	4.615121	9.562053	Pave	Missing	IR1	Lvl	AllPub	Corner	Gtl	NridgHt	Norm	Norm	1Fam	2Story	8	5	1	0	Gable	CompShg	Viny
21	22	45	RM	4.043051	8.915835	Pave	Grvl	Reg	Bnk	AllPub	Inside	Gtl	IDOTRR	Norm	Norm	1Fam	Rare Var	7	7	77	57	Gable	CompShg	Wd S
22	23	20	RL	4.317488	9.184202	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	1Story	8	5	6	6	Hip	CompShg	Viny
23	24	120	RM	3.784190	8.348538	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	MeadowV	Norm	Norm	TwnhsE	1Story	5	7	31	31	Gable	CompShg	Cemr
24	25	20	RL	4.234107	9.017484	Pave	Missing	IR1	Lvl	AllPub	Inside	Gtl	Sawyer	Norm	Norm	1Fam	1Story	5	8	42	9	Gable	CompShg	Plyw
25	26	20	RL	4.700480	9.563108	Pave	Missing	Reg	Lvl	AllPub	Corner	Gtl	NridgHt	Norm	Norm	1Fam	1Story	8	5	2	2	Gable	CompShg	Viny
26	27	20	RL	4.094345	8.881836	Pave	Missing	Reg	Lvl	AllPub	Corner	Gtl	NAmes	Norm	Norm	1Fam	1Story	5	7	59	10	Gable	CompShg	Wd S
27	28	20	RL	4.584967	9.348187	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	NridgHt	Norm	Norm	1Fam	1Story	8	5	3	2	Gable	CompShg	Viny
28	29	20	RL	3.850148	9.700208	Pave	Missing	IR1	Lvl	AllPub	CulDSac	Gtl	NAmes	Norm	Norm	1Fam	1Story	5	6	49	9	Gable	CompShg	Meta
29	30	30	RM	4.094345	8.752107	Pave	Missing	IR1	Lvl	AllPub	Inside	Gtl	BrkSide	Feedr	Rare Var	1Fam	1Story	4	6	81	58	Gable	CompShg	Meta
30	31	70	Rare Var	3.912023	9.047821	Pave	Pave	Reg	Lvl	AllPub	Inside	Gtl	IDOTRR	Feedr	Norm	1Fam	2Story	4	4	88	58	Rare Var	CompShg	BrkF
31	32	20	RL	4.234107	9.052985	Pave	Missing	IR1	Lvl	AllPub	CulDSac	Gtl	Sawyer	Norm	Norm	1Fam	1Story	5	6	42	2	Gable	CompShg	HdBo
32	33	20	RL	4.442651	9.310095	Pave	Missing	Reg	Lvl	AllPub	Corner	Gtl	CollgCr	Norm	Norm	1Fam	1Story	8	5	1	1	Gable	CompShg	Viny
33	34	20	RL	4.248495	9.264071	Pave	Missing	IR1	Lvl	AllPub	Inside	Gtl	NAmes	Norm	Norm	1Fam	1Story	5	5	51	51	Hip	CompShg	BrkF
34	35	120	RL	4.094345	8.897409	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	NridgHt	Norm	Norm	TwnhsE	1Story	9	5	2	2	Hip	CompShg	Meta
35	36	60	RL	4.682131	9.504352	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	NridgHt	Norm	Norm	1Fam	2Story	8	5	2	1	Gable	CompShg	Viny
36	37	20	RL	4.718499	9.292750	Pave	Missing	Reg	Lvl	AllPub	Corner	Gtl	CollgCr	Norm	Norm	1Fam	1Story	5	5	15	14	Gable	CompShg	Viny
37	38	20	RL	4.304065	9.051579	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	NAmes	Norm	Norm	1Fam	1Story	5	6	55	19	Hip	CompShg	Wd S
38	39	20	RL	4.219508	8.977399	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	NAmes	Norm	Norm	1Fam	1Story	5	7	57	3	Gable	CompShg	Viny
39	40	90	RL	4.174387	8.706159	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	Edwards	Norm	Norm	Duplex	1Story	4	5	53	53	Gable	CompShg	AsbS
40	41	20	RL	4.430817	9.066239	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	NAmes	Norm	Norm	1Fam	1Story	6	5	41	41	Gable	CompShg	Wd S
41	42	20	RL	4.744932	9.735365	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	Timber	Norm	Norm	1Fam	1Story	5	6	48	48	Gable	CompShg	Viny
42	43	85	RL	4.234107	9.124782	Pave	Missing	IR1	Lvl	AllPub	CulDSac	Gtl	SawyerW	Norm	Norm	1Fam	SFoyer	5	7	24	24	Gable	CompShg	HdBo
43	44	20	RL	4.234107	9.126959	Pave	Missing	IR1	Lvl	AllPub	CulDSac	Gtl	CollgCr	Norm	Norm	1Fam	1Story	5	6	33	28	Hip	CompShg	Viny
44	45	20	RL	4.248495	8.980298	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	NAmes	Norm	Norm	1Fam	1Story	5	6	47	47	Gable	CompShg	BrkF
45	46	120	RL	4.110874	8.943506	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	NridgHt	Norm	Norm	TwnhsE	1Story	9	5	5	5	Hip	CompShg	Meta
46	47	50	RL	3.871201	9.458918	Pave	Missing	IR1	Lvl	AllPub	CulDSac	Gtl	Mitchel	Norm	Norm	1Fam	1.5Fin	7	5	6	6	Gable	CompShg	Viny
47	48	20	FV	4.430817	9.314340	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	Somerst	Norm	Norm	1Fam	1Story	8	5	1	1	Gable	CompShg	Viny
48	49	190	RM	3.496508	8.402007	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	OldTown	Norm	Norm	2fmCon	2Story	4	5	89	1	Gable	CompShg	Meta
49	50	20	RL	4.189655	8.954415	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	Sawyer	Norm	Norm	1Fam	1Story	5	7	41	41	Gable	CompShg	HdBo
50	51	60	RL	4.234107	9.537411	Pave	Missing	IR2	Lvl	AllPub	Corner	Gtl	Gilbert	Norm	Norm	1Fam	2Story	6	6	10	10	Gable	CompShg	Viny
51	52	50	RM	3.951244	8.738735	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	BrkSide	Norm	Norm	1Fam	1.5Fin	6	6	72	56	Gable	CompShg	Wd S
52	53	90	RM	4.700480	9.044522	Rare Var	Missing	IR2	Bnk	AllPub	Corner	Mod	IDOTRR	Rare Var	Norm	Duplex	1Story	5	5	47	47	Gable	CompShg	Wd S
53	54	20	RL	4.219508	10.825184	Pave	Missing	IR1	Low	AllPub	Inside	Gtl	Rare Var	Norm	Norm	1Fam	1Story	9	5	25	19	Gable	Rare Var	WdSh
54	55	80	RL	4.094345	8.872627	Pave	Missing	Reg	Bnk	AllPub	Inside	Mod	NAmes	Norm	Norm	1Fam	SLvl	5	5	52	52	Gable	CompShg	Meta
55	56	20	RL	4.605170	9.227689	Pave	Missing	IR1	Lvl	AllPub	Inside	Gtl	NAmes	Norm	Norm	1Fam	1Story	6	5	44	44	Gable	CompShg	HdBo
56	57	160	FV	3.178054	7.880426	Pave	Pave	Reg	Lvl	AllPub	Inside	Gtl	Somerst	Norm	Norm	Twnhs	2Story	8	5	10	9	Gable	CompShg	Meta
57	58	60	RL	4.488636	9.362632	Pave	Missing	IR1	Lvl	AllPub	Corner	Gtl	CollgCr	Norm	Norm	1Fam	2Story	7	5	2	2	Gable	CompShg	Viny
58	59	60	RL	4.189655	9.523836	Pave	Missing	IR2	HLS	AllPub	CulDSac	Gtl	StoneBr	Norm	Norm	1Fam	2Story	10	5	0	0	Hip	CompShg	Viny
59	60	20	RL	4.094345	8.881836	Pave	Missing	Reg	Bnk	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	1Story	5	7	36	36	Gable	CompShg	HdBo
60	61	20	RL	4.143135	9.478228	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	SawyerW	Rare Var	Norm	1Fam	1Story	6	5	2	2	Gable	CompShg	Viny
61	62	75	RM	4.094345	8.881836	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	IDOTRR	Norm	Norm	1Fam	Rare Var	5	7	87	11	Gable	CompShg	Meta
62	63	120	RL	3.784190	8.770594	Pave	Missing	IR1	Lvl	AllPub	Inside	Gtl	NridgHt	Norm	Norm	TwnhsE	1Story	8	5	1	1	Gable	CompShg	Viny
63	64	70	RM	3.912023	9.239899	Pave	Missing	IR1	Bnk	AllPub	Inside	Gtl	OldTown	RRAn	Rare Var	1Fam	2Story	7	6	89	60	Gable	CompShg	Stu
64	65	60	RL	4.234107	9.145802	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	2Story	7	5	12	11	Gable	CompShg	Viny
65	66	60	RL	4.330733	9.168580	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	NridgHt	Norm	Norm	1Fam	2Story	8	5	3	2	Gable	CompShg	Viny
66	67	20	RL	4.234107	9.898475	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	NAmes	Rare Var	Norm	1Fam	1Story	7	5	40	21	Gable	CompShg	Plyw
67	68	20	RL	4.276666	9.274723	Pave	Missing	IR1	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	1Story	7	5	4	4	Gable	CompShg	Viny
68	69	30	RM	3.850148	8.435549	Pave	Missing	Reg	Lvl	AllPub	Corner	Gtl	OldTown	Artery	Norm	1Fam	1Story	4	6	65	60	Gable	CompShg	Meta
69	70	50	RL	4.394449	9.654577	Pave	Missing	Reg	Lvl	AllPub	Corner	Gtl	ClearCr	Norm	Norm	1Fam	1.5Fin	7	4	53	53	Gable	CompShg	BrkF
70	71	20	RL	4.553877	9.521568	Pave	Missing	IR1	Lvl	AllPub	Inside	Gtl	NAmes	Norm	Norm	1Fam	1Story	7	6	34	34	Gable	CompShg	Plyw
71	72	20	RL	4.234107	8.935772	Pave	Missing	Reg	Lvl	AllPub	Corner	Gtl	Mitchel	Norm	Norm	1Fam	1Story	4	6	25	1	Gable	CompShg	HdBo
72	73	60	RL	4.304065	9.224342	Pave	Missing	IR1	Lvl	AllPub	Corner	Gtl	Gilbert	Norm	Norm	1Fam	2Story	7	5	11	11	Gable	CompShg	Viny
73	74	20	RL	4.442651	9.230143	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	NAmes	Norm	Norm	1Fam	1Story	5	7	56	7	Gable	CompShg	Wd S
74	75	50	RM	4.094345	8.663888	Pave	Missing	Reg	Lvl	AllPub	Corner	Gtl	OldTown	Norm	Norm	1Fam	2Story	3	6	95	60	Rare Var	CompShg	Viny
75	76	180	RM	3.044522	7.375256	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	MeadowV	Norm	Norm	Twnhs	SLvl	4	5	36	36	Gable	CompShg	Cemr
76	77	20	RL	4.234107	9.044876	Pave	Missing	IR1	Lvl	AllPub	Inside	Gtl	NAmes	Norm	Norm	1Fam	1Story	4	7	52	52	Gable	CompShg	Viny
77	78	50	RM	3.912023	9.063579	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	BrkSide	Norm	Norm	1Fam	1.5Fin	5	5	60	7	Gable	CompShg	Wd S
78	79	90	RL	4.276666	9.285262	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	Sawyer	Norm	Norm	Duplex	1Story	4	5	42	42	Hip	CompShg	HdBo
79	80	50	RM	4.094345	9.253400	Pave	Grvl	Reg	Lvl	AllPub	Corner	Gtl	OldTown	Norm	Norm	1Fam	2Story	5	6	99	28	Gable	CompShg	Wd S
80	81	60	RL	4.605170	9.472705	Pave	Missing	Reg	Lvl	AllPub	Corner	Gtl	NAmes	Norm	Norm	1Fam	2Story	6	6	41	41	Gable	CompShg	Viny
81	82	120	RM	3.465736	8.411833	Pave	Missing	Reg	Lvl	AllPub	FR2	Gtl	Mitchel	Norm	Norm	TwnhsE	1Story	6	5	8	8	Hip	CompShg	Viny
82	83	20	RL	4.356709	9.230731	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	Somerst	Norm	Norm	1Fam	1Story	8	5	1	1	Gable	CompShg	Viny
83	84	20	RL	4.382027	9.092907	Pave	Missing	IR1	Lvl	AllPub	Inside	Gtl	NAmes	Norm	Norm	1Fam	1Story	5	5	47	47	Gable	CompShg	Meta
84	85	80	RL	4.234107	9.051345	Pave	Missing	IR1	Lvl	AllPub	Inside	Gtl	Gilbert	Norm	Norm	1Fam	SLvl	7	5	14	13	Gable	CompShg	HdBo
85	86	60	RL	4.795791	9.684025	Pave	Missing	Reg	Lvl	AllPub	Corner	Gtl	NoRidge	Norm	Norm	1Fam	2Story	8	5	15	14	Hip	CompShg	HdBo
86	87	60	RL	4.804021	9.385218	Pave	Missing	IR2	Lvl	AllPub	Inside	Gtl	Gilbert	Norm	Norm	1Fam	2Story	6	5	4	4	Gable	CompShg	Viny
87	88	160	FV	3.688879	8.281724	Pave	Pave	Reg	Lvl	AllPub	Corner	Gtl	Somerst	Norm	Norm	TwnhsE	2Story	6	5	0	0	Gable	CompShg	Viny
88	89	50	Rare Var	4.653960	9.044286	Pave	Missing	IR1	Lvl	AllPub	Corner	Gtl	IDOTRR	Feedr	Rare Var	1Fam	1.5Fin	3	2	94	27	Hip	CompShg	Plyw
89	90	20	RL	4.094345	8.995909	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	1Story	4	5	13	12	Gable	CompShg	Viny
90	91	20	RL	4.094345	8.881836	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	NAmes	Norm	Norm	1Fam	1Story	4	5	56	56	Gable	CompShg	BrkF
91	92	20	RL	4.442651	9.047821	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	NAmes	Norm	Norm	1Fam	1Story	5	3	45	45	Hip	CompShg	HdBo
92	93	30	RL	4.382027	9.500020	Pave	Grvl	IR1	HLS	AllPub	Inside	Gtl	Crawfor	Norm	Norm	1Fam	1Story	5	7	88	3	Gable	CompShg	Wd S
93	94	190	Rare Var	4.094345	8.881836	Pave	Missing	Reg	Lvl	AllPub	Corner	Gtl	OldTown	Norm	Norm	2fmCon	Rare Var	6	6	97	9	Hip	CompShg	Meta
94	95	60	RL	4.234107	9.141740	Pave	Missing	IR1	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	2Story	6	5	10	10	Gable	CompShg	Viny
95	96	60	RL	4.234107	9.186560	Pave	Missing	IR2	Lvl	AllPub	Corner	Gtl	Gilbert	Norm	Norm	1Fam	2Story	6	8	16	16	Gable	CompShg	Viny
96	97	20	RL	4.356709	9.236398	Pave	Missing	IR1	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	1Story	7	5	7	7	Gable	CompShg	Viny
97	98	20	RL	4.290459	9.298443	Pave	Missing	Reg	HLS	AllPub	Inside	Gtl	Edwards	Norm	Norm	1Fam	1Story	4	5	42	42	Hip	CompShg	HdBo
98	99	30	RL	4.442651	9.270965	Pave	Missing	Reg	Lvl	AllPub	Corner	Gtl	Edwards	Norm	Norm	1Fam	1Story	5	5	90	60	Gable	CompShg	Wd S
99	100	20	RL	4.343805	9.139918	Pave	Missing	IR1	Lvl	AllPub	Inside	Gtl	NAmes	Norm	Norm	1Fam	1Story	4	5	51	51	Gable	CompShg	Plyw

In [34]:
```
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 [36]:
```
dataset.head()
```

Out[36]:

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	Neighborhood	Condition1	Condition2	BldgType	HouseStyle	OverallQual	OverallCond	YearBuilt	YearRemodAdd	RoofStyle	RoofMatl	Exterior1st	E
0	1	60	3	4.174387	9.041922	1	2	0	1	1	0	0	14	2	1	3	5	7	5	5	5	0	0	10	
1	2	20	3	4.382027	9.169518	1	2	0	1	1	2	0	11	1	1	3	3	6	8	31	31	0	0	4	
2	3	60	3	4.219508	9.328123	1	2	1	1	1	0	0	14	2	1	3	5	7	5	7	6	0	0	10	
3	4	70	3	4.094345	9.164296	1	2	1	1	1	1	0	16	2	1	3	5	7	5	91	36	0	0	2	
4	5	60	3	4.430817	9.565214	1	2	1	1	1	2	0	22	2	1	3	5	8	5	8	8	0	0	10	

Feature Scaling

In [37]:
```
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])
```

Out[37]:
```
MinMaxScaler()
```

In [38]:
```
scaler.transform(dataset[feature_scale])
```

Out[38]:
```
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 [39]:
```
data = pd.concat([dataset[['Id', 'SalePrice']].reset_index(drop=True),
                  pd.DataFrame(scaler.transform(dataset[feature_scale]), columns=feature_scale)],
                  axis=1)
```

In [40]:
```
data.head()
```

Out[40]:

	Id	SalePrice	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	Neighborhood	Condition1	Condition2	BldgType	HouseStyle	OverallQual	OverallCond	YearBuilt	YearRemodAdd	RoofStyle	RoofMatl	Ex
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	1.0	0.75	1.0	0.666667	0.500	0.036765	0.098361	0.0	0.0	
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	1.0	0.75	0.6	0.555556	0.875	0.227941	0.524590	0.0	0.0	
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	1.0	0.75	1.0	0.666667	0.500	0.051471	0.114754	0.0	0.0	
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	1.0	0.75	1.0	0.666667	0.500	0.669118	0.606557	0.0	0.0	
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	1.0	0.75	1.0	0.777778	0.500	0.058824	0.147541	0.0	0.0	

In [41]:
```
data.to_csv('x_train_fe.csv', index = False)
```

In [ ]: