In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as mtp
import seaborn as sns
```

In [2]:
```
df=pd.read_csv('C:/Users/aryan/OneDrive/Documents/House Price Dataset.csv')
```

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

Out[3]:
```
(2919, 13)
```

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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2919 entries, 0 to 2918
Data columns (total 13 columns):
 #   Column        Non-Null Count  Dtype  
---  ------        --------------  -----  
 0   Id            2919 non-null   int64  
 1   MSSubClass    2919 non-null   int64  
 2   MSZoning      2915 non-null   object 
 3   LotArea       2919 non-null   int64  
 4   LotConfig     2919 non-null   object 
 5   BldgType      2919 non-null   object 
 6   OverallCond   2919 non-null   int64  
 7   YearBuilt     2919 non-null   int64  
 8   YearRemodAdd  2919 non-null   int64  
 9   Exterior1st   2918 non-null   object 
 10  BsmtFinSF2    2918 non-null   float64
 11  TotalBsmtSF   2918 non-null   float64
 12  SalePrice     1460 non-null   float64
dtypes: float64(3), int64(6), object(4)
memory usage: 296.6+ KB
```

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

Out[5]:
```
Id                0
MSSubClass        0
MSZoning          4
LotArea           0
LotConfig         0
BldgType          0
OverallCond       0
YearBuilt         0
YearRemodAdd      0
Exterior1st       1
BsmtFinSF2        1
TotalBsmtSF       1
SalePrice      1459
dtype: int64
```

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

Out[6]:

	Id	MSSubClass	MSZoning	LotArea	LotConfig	BldgType	OverallCond	YearBuilt	YearRemod
0	0	60	RL	8450	Inside	1Fam	5	2003	2
1	1	20	RL	9600	FR2	1Fam	8	1976	1
2	2	60	RL	11250	Inside	1Fam	5	2001	2
3	3	70	RL	9550	Corner	1Fam	5	1915	1
4	4	60	RL	14260	FR2	1Fam	5	2000	2

In [7]:
```
df.describe()
```

Out[7]:

	Id	MSSubClass	LotArea	OverallCond	YearBuilt	YearRemodAdd	BsmtI
count	2919.000000	2919.000000	2919.000000	2919.000000	2919.000000	2919.000000	2918.0
mean	1459.000000	57.137718	10168.114080	5.564577	1971.312778	1984.264474	49.5
std	842.787043	42.517628	7886.996359	1.113131	30.291442	20.894344	169.2
min	0.000000	20.000000	1300.000000	1.000000	1872.000000	1950.000000	0.0
25%	729.500000	20.000000	7478.000000	5.000000	1953.500000	1965.000000	0.0
50%	1459.000000	50.000000	9453.000000	5.000000	1973.000000	1993.000000	0.0
75%	2188.500000	70.000000	11570.000000	6.000000	2001.000000	2004.000000	0.0
max	2918.000000	190.000000	215245.000000	9.000000	2010.000000	2010.000000	1526.0

In [8]:
```
nulldf= df[df['SalePrice'].isnull()]
```

In [9]:
```
nulldf.head()
```

Out[9]:

	Id	MSSubClass	MSZoning	LotArea	LotConfig	BldgType	OverallCond	YearBuilt	YearF
1460	1460	20	RH	11622	Inside	1Fam	6	1961	
1461	1461	20	RL	14267	Corner	1Fam	6	1958	
1462	1462	60	RL	13830	Inside	1Fam	5	1997	
1463	1463	60	RL	9978	Inside	1Fam	6	1998	
1464	1464	120	RL	5005	Inside	TwnhsE	5	1992	

In [10]:
```
df['SalePrice']= df['SalePrice'].fillna(df['SalePrice'].mean())
```

In [11]:
```
df.isnull().sum()
```

Out[11]:
```
Id              0
MSSubClass      0
MSZoning        4
LotArea         0
LotConfig       0
BldgType        0
OverallCond     0
YearBuilt       0
YearRemodAdd    0
Exterior1st     1
BsmtFinSF2      1
TotalBsmtSF     1
SalePrice       0
dtype: int64
```

In [12]:
```
df['MSZoning'].value_counts()
```

Out[12]:
```
RL         2265
RM          460
FV          139
RH           26
C (all)      25
Name: MSZoning, dtype: int64
```

In [13]:
```
df['MSZoning']= df['MSZoning'].fillna('NA')
```

In [14]:
```
df.isnull().sum()
```

Out[14]:
```
Id              0
MSSubClass      0
MSZoning        0
LotArea         0
LotConfig       0
BldgType        0
OverallCond     0
YearBuilt       0
YearRemodAdd    0
Exterior1st     1
BsmtFinSF2      1
TotalBsmtSF     1
SalePrice       0
dtype: int64
```

In [15]:
```
df['Exterior1st'].value_counts()
```

Out[15]:
```
VinylSd    1025
MetalSd     450
HdBoard     442
Wd Sdng     411
Plywood     221
CemntBd     126
BrkFace      87
WdShing      56
AsbShng      44
Stucco       43
BrkComm       6
AsphShn       2
Stone         2
CBlock        2
ImStucc       1
Name: Exterior1st, dtype: int64
```

In [16]:
```
df['Exterior1st']= df['Exterior1st'].fillna('NA')
```

In [17]:
```
df.isnull().sum()
```

Out[17]:
```
Id              0
MSSubClass      0
MSZoning        0
LotArea         0
LotConfig       0
BldgType        0
OverallCond     0
YearBuilt       0
YearRemodAdd    0
Exterior1st     0
BsmtFinSF2      1
TotalBsmtSF     1
SalePrice       0
dtype: int64
```

In [18]:
```
df['BsmtFinSF2'].value_counts()
```

Out[18]:
```
0.0       2571
180.0        5
294.0        5
435.0        3
483.0        3
          ...
600.0        1
211.0        1
1031.0       1
438.0        1
297.0        1
Name: BsmtFinSF2, Length: 272, dtype: int64
```

In [19]:
```
df['BsmtFinSF2']= df['BsmtFinSF2'].fillna(df['BsmtFinSF2'].mean())
```

In [20]:
```
df['TotalBsmtSF'].value_counts()
```

Out[20]:
```
0.0       78
864.0     74
672.0     29
912.0     26
1040.0    25
          ..
1571.0     1
2633.0     1
757.0      1
873.0      1
1381.0     1
Name: TotalBsmtSF, Length: 1058, dtype: int64
```

In [21]:
```
df['TotalBsmtSF']= df['TotalBsmtSF'].fillna(df['TotalBsmtSF'].mean())
```

In [22]:
```
df.isnull().sum()
```

Out[22]:
```
Id              0
MSSubClass      0
MSZoning        0
LotArea         0
LotConfig       0
BldgType        0
OverallCond     0
YearBuilt       0
YearRemodAdd    0
Exterior1st     0
BsmtFinSF2      0
TotalBsmtSF     0
SalePrice       0
dtype: int64
```

In [23]:
```
df.describe()
```

Out[23]:

	Id	MSSubClass	LotArea	OverallCond	YearBuilt	YearRemodAdd	BsmtI
count	2919.000000	2919.000000	2919.000000	2919.000000	2919.000000	2919.000000	2919.0
mean	1459.000000	57.137718	10168.114080	5.564577	1971.312778	1984.264474	49.5
std	842.787043	42.517628	7886.996359	1.113131	30.291442	20.894344	169.1
min	0.000000	20.000000	1300.000000	1.000000	1872.000000	1950.000000	0.0
25%	729.500000	20.000000	7478.000000	5.000000	1953.500000	1965.000000	0.0
50%	1459.000000	50.000000	9453.000000	5.000000	1973.000000	1993.000000	0.0
75%	2188.500000	70.000000	11570.000000	6.000000	2001.000000	2004.000000	0.0
max	2918.000000	190.000000	215245.000000	9.000000	2010.000000	2010.000000	1526.0

In [24]:
```
df.head(25)
```

Out[24]:

	Id	MSSubClass	MSZoning	LotArea	LotConfig	BldgType	OverallCond	YearBuilt	YearRemo
0	0	60	RL	8450	Inside	1Fam	5	2003	
1	1	20	RL	9600	FR2	1Fam	8	1976	
2	2	60	RL	11250	Inside	1Fam	5	2001	
3	3	70	RL	9550	Corner	1Fam	5	1915	
4	4	60	RL	14260	FR2	1Fam	5	2000	
5	5	50	RL	14115	Inside	1Fam	5	1993	
6	6	20	RL	10084	Inside	1Fam	5	2004	
7	7	60	RL	10382	Corner	1Fam	6	1973	
8	8	50	RM	6120	Inside	1Fam	5	1931	
9	9	190	RL	7420	Corner	2fmCon	6	1939	
10	10	20	RL	11200	Inside	1Fam	5	1965	
11	11	60	RL	11924	Inside	1Fam	5	2005	
12	12	20	RL	12968	Inside	1Fam	6	1962	
13	13	20	RL	10652	Inside	1Fam	5	2006	
14	14	20	RL	10920	Corner	1Fam	5	1960	
15	15	45	RM	6120	Corner	1Fam	8	1929	
16	16	20	RL	11241	CulDSac	1Fam	7	1970	
17	17	90	RL	10791	Inside	Duplex	5	1967	
18	18	20	RL	13695	Inside	1Fam	5	2004	
19	19	20	RL	7560	Inside	1Fam	6	1958	
20	20	60	RL	14215	Corner	1Fam	5	2005	
21	21	45	RM	7449	Inside	1Fam	7	1930	
22	22	20	RL	9742	Inside	1Fam	5	2002	
23	23	120	RM	4224	Inside	TwnhsE	7	1976	
24	24	20	RL	8246	Inside	1Fam	8	1968	

In [25]:
```
df['LotConfig'].value_counts()
```

Out[25]:
```
Inside     2133
Corner      511
CulDSac     176
FR2          85
FR3          14
Name: LotConfig, dtype: int64
```

In [26]:
```
df['Exterior1st'].value_counts()
```

Out[26]:
```
VinylSd    1025
MetalSd     450
HdBoard     442
Wd Sdng     411
Plywood     221
CemntBd     126
BrkFace      87
WdShing      56
AsbShng      44
Stucco       43
BrkComm       6
AsphShn       2
Stone         2
CBlock        2
ImStucc       1
NA            1
Name: Exterior1st, dtype: int64
```

In [27]:
```
df.groupby('Exterior1st', as_index=False).sum()
```

Out[27]:

	Exterior1st	Id	MSSubClass	LotArea	OverallCond	YearBuilt	YearRemodAdd	BsmtFi
0	AsbShng	69689	2720	377200	219	84874	86306	858.0
1	AsphShn	3146	280	20145	8	3880	3915	0.0
2	BrkComm	10636	230	69158	31	11700	11828	507.0
3	BrkFace	123358	4055	1185874	482	170244	171194	13846.0
4	CBlock	4274	70	36650	9	3871	3901	105.0
5	CemntBd	191494	11340	1074435	689	250532	251298	2595.0
6	HdBoard	644651	25020	4480124	2487	872203	874897	31612.0
7	ImStucc	1187	20	12461	5	1994	1995	0.0
8	MetalSd	666007	28920	3856174	2616	880423	888927	13951.5
9	NA	2151	30	19550	7	1940	2007	0.0
10	Plywood	325276	14510	2780250	1235	435931	437247	24265.0
11	Stone	2176	40	29613	13	3932	3980	947.0
12	Stucco	56957	2195	419382	264	82950	84809	2015.0
13	VinylSd	1475970	54445	10260705	5425	2042643	2048756	23344.0
14	Wd Sdng	598395	20320	4504553	2433	797613	810443	27358.0
15	WdShing	83454	2590	554451	320	109532	110565	3327.0

In [28]:
```
df['SalePrice'].describe()
```

Out[28]:
```
count      2919.000000
mean     180921.195890
std       56174.332503
min       34900.000000
25%      163000.000000
50%      180921.195890
75%      180921.195890
max      755000.000000
Name: SalePrice, dtype: float64
```

In [29]:
```
print(df.shape)
df
```

```
(2919, 13)
```

Out[29]:

	Id	MSSubClass	MSZoning	LotArea	LotConfig	BldgType	OverallCond	YearBuilt	YearR
0	0	60	RL	8450	Inside	1Fam	5	2003	
1	1	20	RL	9600	FR2	1Fam	8	1976	
2	2	60	RL	11250	Inside	1Fam	5	2001	
3	3	70	RL	9550	Corner	1Fam	5	1915	
4	4	60	RL	14260	FR2	1Fam	5	2000	
...	...	...	...	...	...	...	...	...	...
2914	2914	160	RM	1936	Inside	Twnhs	7	1970	
2915	2915	160	RM	1894	Inside	TwnhsE	5	1970	
2916	2916	20	RL	20000	Inside	1Fam	7	1960	
2917	2917	85	RL	10441	Inside	1Fam	5	1992	
2918	2918	60	RL	9627	Inside	1Fam	5	1993	

2919 rows × 13 columns

In [30]:
```
x= df['YearBuilt']
y= df['SalePrice']
print(x.shape)
print(y.shape)
```

```
(2919,)
(2919,)
```

In [31]:
```
from sklearn.model_selection import train_test_split
xtrain,xtest,ytrain,ytest = train_test_split(x,y,test_size=0.2)
print(xtrain.shape)
print(xtest.shape)
print(ytrain.shape)
print(ytest.shape)
```

```
(2335,)
(584,)
(2335,)
(584,)
```

In [32]:
```
import statsmodels.api as sm
model= sm.OLS(ytrain, xtrain).fit()
```

In [33]:
```
print(model.summary())
```

OLS Regression Results

Dep. Variable:	SalePrice	R-squared (uncentered):	0.911
Model:	OLS	Adj. R-squared (uncentered):	0.911
Method:	Least Squares	F-statistic:	2.39e+04
Date:	Mon, 27 Mar 2023	Prob (F-statistic):	0.00
Time:	18:10:00	Log-Likelihood:	-28867.
No. Observations:	2335	AIC:	5.774e+04
Df Residuals:	2334	BIC:	5.774e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
YearBuilt	91.9276	0.594	154.677	0.000	90.762	93.093

Omnibus:	1580.760	Durbin-Watson:	1.938
Prob(Omnibus):	0.000	Jarque-Bera (JB):	34382.543
Skew:	2.884	Prob(JB):	0.00
Kurtosis:	20.892	Cond. No.	1.00

Notes:
[1] R^2 is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.