Project 7 - Feature EngineeringFeature Engineering

December 22, 2022

1 Perform the following steps:

1.1 1. Understand the dataset:

a. Identify the shape of the dataset

b. Identify variables with null values

c. Identify variables with unique values

1.2 2. Generate a separate dataset for numerical and categorical variables

1.3 3. EDA of numerical variables:

a. Missing value treatment

b. Identify the skewness and distribution

c. Identify significant variables using a correlation matrix

d. Pair plot for distribution and density

1.4 4. EDA of categorical variables

a. Missing value treatment

b. Count plot and box plot for bivariate analysis

c. Identify significant variables using p-values and Chi-Square values

1.5 5. Combine all the significant categorical and numerical variables

1.6 6. Plot box plot for the new dataset to find the variables with outliers

Note: The last two points are performed to make the new dataset ready for training and prediction.

[1]:
```
#import the librabries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

[2]:
```
df = pd.read_csv('PEP1.csv')
```

[3]:
```
df
```

[3]:
```
         Id  MSSubClass MSZoning  LotFrontage  LotArea Street Alley LotShape  \
0         1          60       RL         65.0     8450   Pave   NaN      Reg
1         2          20       RL         80.0     9600   Pave   NaN      Reg
2         3          60       RL         68.0    11250   Pave   NaN      IR1
3         4          70       RL         60.0     9550   Pave   NaN      IR1
4         5          60       RL         84.0    14260   Pave   NaN      IR1
…       …          …       …           …       …    …   …       …
1455   1456          60       RL         62.0     7917   Pave   NaN      Reg
1456   1457          20       RL         85.0    13175   Pave   NaN      Reg
1457   1458          70       RL         66.0     9042   Pave   NaN      Reg
1458   1459          20       RL         68.0     9717   Pave   NaN      Reg
1459   1460          20       RL         75.0     9937   Pave   NaN      Reg

     LandContour Utilities  … PoolArea PoolQC  Fence MiscFeature MiscVal  \
0            Lvl    AllPub  …        0    NaN    NaN         NaN       0
1            Lvl    AllPub  …        0    NaN    NaN         NaN       0
2            Lvl    AllPub  …        0    NaN    NaN         NaN       0
3            Lvl    AllPub  …        0    NaN    NaN         NaN       0
4            Lvl    AllPub  …        0    NaN    NaN         NaN       0
…          …       …  …       …     …    …           …      …
1455         Lvl    AllPub  …        0    NaN    NaN         NaN       0
1456         Lvl    AllPub  …        0    NaN  MnPrv         NaN       0
1457         Lvl    AllPub  …        0    NaN  GdPrv        Shed    2500
1458         Lvl    AllPub  …        0    NaN    NaN         NaN       0
1459         Lvl    AllPub  …        0    NaN    NaN         NaN       0

     MoSold YrSold  SaleType  SaleCondition  SalePrice
0         2   2008        WD         Normal     208500
1         5   2007        WD         Normal     181500
2         9   2008        WD         Normal     223500
3         2   2006        WD        Abnorml     140000
4        12   2008        WD         Normal     250000
…      …     …       …             …        …
1455      8   2007        WD         Normal     175000
1456      2   2010        WD         Normal     210000
1457      5   2010        WD         Normal     266500
1458      4   2010        WD         Normal     142125
1459      6   2008        WD         Normal     147500

[1460 rows x 81 columns]
```

[4]:
```
df.columns
```

[4]:
```
Index(['Id', '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', 'GrLivArea', 'BsmtFullBath', 'BsmtHalfBath', 'FullBath',
       'HalfBath', 'BedroomAbvGr', 'KitchebvGr', 'KitchenQual', 'TotRmsAbvGrd',
       'Functiol', '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'],
      dtype='object')
```

2 1. Understand the dataset:

a. Identify the shape of the dataset

b. Identify variables with null values

c. Identify variables with unique values

2.1 1. a. Identify the shape of the dataset

[5]:
```
df.shape
```

[5]:
```
(1460, 81)
```

2.2 1. b. Identify variables with null values

[6]:
```
df.isnull().sum().any()
```

[6]:
```
True
```

[7]:
```
df.isnull().sum().to_frame()
```

[7]:

	0
Id	0
MSSubClass	0
MSZoning	0
LotFrontage	259
LotArea	0
…	…
MoSold	0
YrSold	0
SaleType	0

```
SaleCondition    0
SalePrice        0

[81 rows x 1 columns]
```

2.3 1. c. Identify variables with unique values

[8]:
```
df.nunique().to_frame()
```

[8]:
```
                  0
Id             1460
MSSubClass       15
MSZoning          5
LotFrontage     110
LotArea        1073
…                 …
MoSold           12
YrSold            5
SaleType          9
SaleCondition     6
SalePrice       663

[81 rows x 1 columns]
```

3 2. Generate a separate dataset for numerical and categorical variables

[9]:
```
df.dtypes
```

[9]:
```
Id                int64
MSSubClass        int64
MSZoning         object
LotFrontage     float64
LotArea           int64
                  …
MoSold            int64
YrSold            int64
SaleType         object
SaleCondition    object
SalePrice         int64
Length: 81, dtype: object
```

[10]:
```
df_num=df.select_dtypes(exclude='object')
df_num.dtypes
```

[10]:
```
Id               int64
MSSubClass       int64
LotFrontage    float64
LotArea          int64
OverallQual      int64
OverallCond      int64
YearBuilt        int64
YearRemodAdd     int64
MasVnrArea     float64
BsmtFinSF1       int64
BsmtFinSF2       int64
BsmtUnfSF        int64
TotalBsmtSF      int64
1stFlrSF         int64
2ndFlrSF         int64
LowQualFinSF     int64
GrLivArea        int64
BsmtFullBath     int64
BsmtHalfBath     int64
FullBath         int64
HalfBath         int64
BedroomAbvGr     int64
KitchebvGr       int64
TotRmsAbvGrd     int64
Fireplaces       int64
GarageYrBlt    float64
GarageCars       int64
GarageArea       int64
WoodDeckSF       int64
OpenPorchSF      int64
EnclosedPorch    int64
3SsnPorch        int64
ScreenPorch      int64
PoolArea         int64
MiscVal          int64
MoSold           int64
YrSold           int64
SalePrice        int64
dtype: object
```

[11]:
```
df_num.head()
```

[11]:

	Id	MSSubClass	LotFrontage	LotArea	OverallQual	OverallCond	YearBuilt	\
0	1	60	65.0	8450	7	5	2003	
1	2	20	80.0	9600	6	8	1976	
2	3	60	68.0	11250	7	5	2001	
3	4	70	60.0	9550	7	5	1915	

```
4   5          60         84.0    14260            8            5       2000

   YearRemodAdd  MasVnrArea  BsmtFinSF1  …  WoodDeckSF  OpenPorchSF  \
0          2003       196.0         706  …           0           61
1          1976         0.0         978  …         298            0
2          2002       162.0         486  …           0           42
3          1970         0.0         216  …           0           35
4          2000       350.0         655  …         192           84

   EnclosedPorch  3SsnPorch  ScreenPorch  PoolArea  MiscVal  MoSold  YrSold  \
0              0          0            0         0        0       2    2008
1              0          0            0         0        0       5    2007
2              0          0            0         0        0       9    2008
3            272          0            0         0        0       2    2006
4              0          0            0         0        0      12    2008

   SalePrice
0     208500
1     181500
2     223500
3     140000
4     250000

[5 rows x 38 columns]
```

[12]:
```
df_num.shape
```

[12]:
```
(1460, 38)
```

[13]:
```
df_obj=df.select_dtypes(exclude=['float','int'])
df_obj.dtypes
```

[13]:
```
MSZoning         object
Street           object
Alley            object
LotShape         object
LandContour      object
Utilities        object
LotConfig        object
LandSlope        object
Neighborhood     object
Condition1       object
Condition2       object
BldgType         object
HouseStyle       object
RoofStyle        object
RoofMatl         object
```

```
Exterior1st      object
Exterior2nd      object
MasVnrType       object
ExterQual        object
ExterCond        object
Foundation       object
BsmtQual         object
BsmtCond         object
BsmtExposure     object
BsmtFinType1     object
BsmtFinType2     object
Heating          object
HeatingQC        object
CentralAir       object
Electrical       object
KitchenQual      object
Functiol         object
FireplaceQu      object
GarageType       object
GarageFinish     object
GarageQual       object
GarageCond       object
PavedDrive       object
PoolQC           object
Fence            object
MiscFeature      object
SaleType         object
SaleCondition    object
dtype: object
```

[14]:
```
df_obj.head()
```

[14]:

	MSZoning	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	\
0	RL	Pave	NaN	Reg	Lvl	AllPub	Inside	Gtl	
1	RL	Pave	NaN	Reg	Lvl	AllPub	FR2	Gtl	
2	RL	Pave	NaN	IR1	Lvl	AllPub	Inside	Gtl	
3	RL	Pave	NaN	IR1	Lvl	AllPub	Corner	Gtl	
4	RL	Pave	NaN	IR1	Lvl	AllPub	FR2	Gtl	

	Neighborhood	Condition1	…	GarageType	GarageFinish	GarageQual	GarageCond	\
0	CollgCr	Norm	…	Attchd	RFn	TA	TA	
1	Veenker	Feedr	…	Attchd	RFn	TA	TA	
2	CollgCr	Norm	…	Attchd	RFn	TA	TA	
3	Crawfor	Norm	…	Detchd	Unf	TA	TA	
4	NoRidge	Norm	…	Attchd	RFn	TA	TA	

	PavedDrive	PoolQC	Fence	MiscFeature	SaleType	SaleCondition

```
0          Y    NaN   NaN         NaN       WD        Normal
1          Y    NaN   NaN         NaN       WD        Normal
2          Y    NaN   NaN         NaN       WD        Normal
3          Y    NaN   NaN         NaN       WD       Abnorml
4          Y    NaN   NaN         NaN       WD        Normal

[5 rows x 43 columns]
```

[15]:
```
df_obj.shape
```

[15]:
```
(1460, 43)
```

4 3. EDA of numerical variables:

a. Missing value treatment

b. Identify the skewness and distribution

c. Identify significant variables using a correlation matrix

d. Pair plot for distribution and density

4.1 3. a. Missing value treatment

[16]:
```
df_num.isnull().sum()[df_num.isnull().sum()>0]
```

[16]:
```
LotFrontage    259
MasVnrArea       8
GarageYrBlt     81
dtype: int64
```

[17]:
```
(df_num.isnull().sum()/len(df_num))*100
```

[17]:
```
Id               0.000000
MSSubClass       0.000000
LotFrontage     17.739726
LotArea          0.000000
OverallQual      0.000000
OverallCond      0.000000
YearBuilt        0.000000
YearRemodAdd     0.000000
MasVnrArea       0.547945
BsmtFinSF1       0.000000
BsmtFinSF2       0.000000
BsmtUnfSF        0.000000
TotalBsmtSF      0.000000
1stFlrSF         0.000000
2ndFlrSF         0.000000
LowQualFinSF     0.000000
```

```
GrLivArea         0.000000
BsmtFullBath      0.000000
BsmtHalfBath      0.000000
FullBath          0.000000
HalfBath          0.000000
BedroomAbvGr      0.000000
KitchebvGr        0.000000
TotRmsAbvGrd      0.000000
Fireplaces        0.000000
GarageYrBlt       5.547945
GarageCars        0.000000
GarageArea        0.000000
WoodDeckSF        0.000000
OpenPorchSF       0.000000
EnclosedPorch     0.000000
3SsnPorch         0.000000
ScreenPorch       0.000000
PoolArea          0.000000
MiscVal           0.000000
MoSold            0.000000
YrSold            0.000000
SalePrice         0.000000
dtype: float64
```

[18]:
```
df_num.dropna(inplace=True)
```

[19]:
```
df_num.shape
```

[19]:
```
(1121, 38)
```

4.2 3. b. Identify the skewness and distribution

[20]:
```
skew = df_num.skew(axis=0)
skew
```

[20]:
```
Id                0.018663
MSSubClass        1.412907
LotFrontage       2.251197
LotArea          15.608113
OverallQual       0.287800
OverallCond       0.846451
YearBuilt        -0.618350
YearRemodAdd     -0.565757
MasVnrArea        2.706945
BsmtFinSF1        1.934077
BsmtFinSF2        4.399358
BsmtUnfSF         0.875774
```

```
TotalBsmtSF       1.754916
1stFlrSF          1.363783
2ndFlrSF          0.807411
LowQualFinSF     10.020823
GrLivArea         1.549961
BsmtFullBath      0.568804
BsmtHalfBath      4.107874
FullBath          0.015822
HalfBath          0.638178
BedroomAbvGr      0.074427
KitchebvGr        4.822542
TotRmsAbvGrd      0.723117
Fireplaces        0.643698
GarageYrBlt      -0.641738
GarageCars        0.206017
GarageArea        0.733894
WoodDeckSF        1.549793
OpenPorchSF       2.403928
EnclosedPorch     3.173250
3SsnPorch        10.854868
ScreenPorch       4.019111
PoolArea         13.783823
MiscVal           9.699989
MoSold            0.173039
YrSold            0.106730
SalePrice         1.933615
dtype: float64
```

[21]:
```
def func(num):
    if num>0.5:
        return 'Positive Skew'
    elif num<-0.5:
        return 'Negative Skew'
    else:
        return 'Normal Distribution'
```

[22]:
```
skew.apply(func)
```

[22]:
```
Id                Normal Distribution
MSSubClass              Positive Skew
LotFrontage             Positive Skew
LotArea                 Positive Skew
OverallQual       Normal Distribution
OverallCond             Positive Skew
YearBuilt               Negative Skew
YearRemodAdd            Negative Skew
MasVnrArea              Positive Skew
```

```
BsmtFinSF1             Positive Skew
BsmtFinSF2             Positive Skew
BsmtUnfSF              Positive Skew
TotalBsmtSF            Positive Skew
1stFlrSF               Positive Skew
2ndFlrSF               Positive Skew
LowQualFinSF           Positive Skew
GrLivArea              Positive Skew
BsmtFullBath           Positive Skew
BsmtHalfBath           Positive Skew
FullBath         Normal Distribution
HalfBath               Positive Skew
BedroomAbvGr     Normal Distribution
KitchebvGr             Positive Skew
TotRmsAbvGrd           Positive Skew
Fireplaces             Positive Skew
GarageYrBlt            Negative Skew
GarageCars       Normal Distribution
GarageArea             Positive Skew
WoodDeckSF             Positive Skew
OpenPorchSF            Positive Skew
EnclosedPorch          Positive Skew
3SsnPorch              Positive Skew
ScreenPorch            Positive Skew
PoolArea               Positive Skew
MiscVal                Positive Skew
MoSold           Normal Distribution
YrSold           Normal Distribution
SalePrice              Positive Skew
dtype: object
```

4.3 3. c. Identify significant variables using a correlation matrix

[23]:
```
df_num.corr()
```

[23]:
```
                    Id  MSSubClass  LotFrontage   LotArea  OverallQual  \
Id            1.000000    0.021937    -0.013289 -0.040711    -0.058269
MSSubClass    0.021937    1.000000    -0.386940 -0.198096     0.029522
LotFrontage  -0.013289   -0.386940     1.000000  0.421184     0.241322
LotArea      -0.040711   -0.198096     0.421184  1.000000     0.167525
OverallQual  -0.058269    0.029522     0.241322  0.167525     1.000000
OverallCond   0.004387   -0.087859    -0.046312 -0.034348    -0.163157
YearBuilt    -0.020862    0.025800     0.109726  0.029205     0.589385
YearRemodAdd -0.027664    0.006645     0.086414  0.026848     0.570757
MasVnrArea   -0.073472    0.040240     0.189969  0.106115     0.423988
BsmtFinSF1   -0.013751   -0.070389     0.241352  0.230441     0.249500
BsmtFinSF2    0.012544   -0.075439     0.049305  0.138234    -0.068506
```

BsmtUnfSF	-0.012985	-0.145582	0.115306	0.011288	0.322663
TotalBsmtSF	-0.023129	-0.247781	0.387620	0.302554	0.563960
1stFlrSF	-0.008046	-0.252249	0.451085	0.329679	0.514453
2ndFlrSF	-0.002346	0.319328	0.075004	0.074612	0.273197
LowQualFinSF	-0.039933	0.024704	0.011148	0.020039	-0.008118
GrLivArea	-0.011068	0.083365	0.396306	0.307164	0.607466
BsmtFullBath	0.026113	-0.014681	0.118088	0.179052	0.126834
BsmtHalfBath	-0.026774	0.012310	0.000434	-0.014282	-0.053283
FullBath	0.007220	0.131278	0.185785	0.129073	0.576875
HalfBath	-0.010409	0.203971	0.045678	0.045183	0.251690
BedroomAbvGr	0.039831	-0.032971	0.270404	0.137269	0.094882
KitchebvGr	0.025913	0.266012	-0.003546	-0.018942	-0.178735
TotRmsAbvGrd	0.020012	0.047209	0.348421	0.237918	0.451008
Fireplaces	-0.018273	-0.031122	0.260321	0.255755	0.415294
GarageYrBlt	-0.002039	0.054701	0.069878	0.013731	0.560425
GarageCars	-0.008125	-0.027411	0.286587	0.172428	0.593803
GarageArea	-0.025889	-0.092607	0.356851	0.211362	0.550659
WoodDeckSF	-0.025060	-0.017988	0.082166	0.133576	0.282512
OpenPorchSF	-0.001972	0.004054	0.161815	0.099170	0.340679
EnclosedPorch	0.009935	-0.017790	0.014261	-0.023631	-0.144344
3SsnPorch	-0.066833	-0.039739	0.069716	0.012520	0.017331
ScreenPorch	0.015183	-0.021789	0.035906	0.072517	0.055296
PoolArea	0.048010	0.003166	0.211746	0.109147	0.080131
MiscVal	0.045799	-0.040689	0.001471	0.012790	-0.062064
MoSold	-0.000570	-0.027170	0.018815	0.008998	0.079895
YrSold	0.013407	-0.012448	0.013267	-0.006904	-0.008903
SalePrice	-0.047122	-0.088032	0.344270	0.299962	0.797881

	OverallCond	YearBuilt	YearRemodAdd	MasVnrArea	BsmtFinSF1	\
Id	0.004387	-0.020862	-0.027664	-0.073472	-0.013751	
MSSubClass	-0.087859	0.025800	0.006645	0.040240	-0.070389	
LotFrontage	-0.046312	0.109726	0.086414	0.189969	0.241352	
LotArea	-0.034348	0.029205	0.026848	0.106115	0.230441	
OverallQual	-0.163157	0.589385	0.570757	0.423988	0.249500	
OverallCond	1.000000	-0.426462	0.039402	-0.166762	-0.054788	
YearBuilt	-0.426462	1.000000	0.623171	0.332190	0.236941	
YearRemodAdd	0.039402	0.623171	1.000000	0.193376	0.120774	
MasVnrArea	-0.166762	0.332190	0.193376	1.000000	0.285331	
BsmtFinSF1	-0.054788	0.236941	0.120774	0.285331	1.000000	
BsmtFinSF2	0.042314	-0.054414	-0.057024	-0.075261	-0.035780	
BsmtUnfSF	-0.148630	0.177545	0.199893	0.110067	-0.502225	
TotalBsmtSF	-0.192762	0.409134	0.308696	0.384434	0.530917	
1stFlrSF	-0.164251	0.308875	0.281436	0.363209	0.468020	
2ndFlrSF	0.005985	-0.011621	0.103627	0.180567	-0.120823	
LowQualFinSF	0.048720	-0.164359	-0.053479	-0.062930	-0.050824	
GrLivArea	-0.112231	0.204967	0.290050	0.414024	0.239888	
BsmtFullBath	-0.060943	0.182800	0.111897	0.110379	0.651727	

BsmtHalfBath	0.122960	-0.049645	-0.017049	-0.007035	0.061963
FullBath	-0.229848	0.500495	0.467563	0.285561	0.052313
HalfBath	-0.079023	0.220000	0.164203	0.195273	0.007545
BedroomAbvGr	0.004643	-0.061580	-0.075812	0.114310	-0.104275
KitchebvGr	-0.092644	-0.171920	-0.181803	-0.023647	-0.062920
TotRmsAbvGrd	-0.096901	0.121417	0.181995	0.315604	0.080207
Fireplaces	-0.022290	0.133077	0.125898	0.252525	0.270306
GarageYrBlt	-0.343206	0.823520	0.645808	0.277095	0.160356
GarageCars	-0.267859	0.532563	0.462663	0.375269	0.196443
GarageArea	-0.226347	0.471286	0.407471	0.382162	0.286657
WoodDeckSF	-0.010835	0.238548	0.244602	0.174649	0.206246
OpenPorchSF	-0.076273	0.235432	0.260521	0.129532	0.127900
EnclosedPorch	0.062748	-0.392693	-0.214115	-0.116832	-0.105410
3SsnPorch	-0.006861	0.027948	0.026304	0.022331	0.021831
ScreenPorch	0.087030	-0.063694	-0.034288	0.052646	0.059635
PoolArea	-0.023566	0.006717	0.019307	0.021648	0.194349
MiscVal	0.119772	-0.096973	-0.040420	-0.054044	0.003027
MoSold	-0.014236	0.013784	0.026884	0.015850	-0.015281
YrSold	0.041003	-0.004585	0.041302	-0.017569	0.010224
SalePrice	-0.124391	0.525394	0.521253	0.488658	0.390301

	…	WoodDeckSF	OpenPorchSF	EnclosedPorch	3SsnPorch	\
Id	…	-0.025060	-0.001972	0.009935	-0.066833	
MSSubClass	…	-0.017988	0.004054	-0.017790	-0.039739	
LotFrontage	…	0.082166	0.161815	0.014261	0.069716	
LotArea	…	0.133576	0.099170	-0.023631	0.012520	
OverallQual	…	0.282512	0.340679	-0.144344	0.017331	
OverallCond	…	-0.010835	-0.076273	0.062748	-0.006861	
YearBuilt	…	0.238548	0.235432	-0.392693	0.027948	
YearRemodAdd	…	0.244602	0.260521	-0.214115	0.026304	
MasVnrArea	…	0.174649	0.129532	-0.116832	0.022331	
BsmtFinSF1	…	0.206246	0.127900	-0.105410	0.021831	
BsmtFinSF2	…	0.032338	0.010518	0.047221	-0.030848	
BsmtUnfSF	…	0.005391	0.151572	-0.035791	0.021502	
TotalBsmtSF	…	0.233664	0.291286	-0.130223	0.033743	
1stFlrSF	…	0.237628	0.244846	-0.113595	0.037505	
2ndFlrSF	…	0.114480	0.203460	0.076479	-0.027471	
LowQualFinSF	…	-0.017374	0.032968	0.060988	0.002171	
GrLivArea	…	0.269703	0.353534	-0.014874	0.004823	
BsmtFullBath	…	0.157510	0.081623	-0.042631	-0.007893	
BsmtHalfBath	…	0.054066	-0.060347	0.000854	0.056402	
FullBath	…	0.215028	0.286248	-0.164548	0.032051	
HalfBath	…	0.114153	0.194016	-0.080586	-0.002422	
BedroomAbvGr	…	0.077918	0.079124	0.040681	-0.029136	
KitchebvGr	…	-0.099832	-0.060133	0.013411	-0.023299	
TotRmsAbvGrd	…	0.190527	0.246714	-0.031651	-0.023904	
Fireplaces	…	0.177763	0.185274	-0.034478	-0.001002	

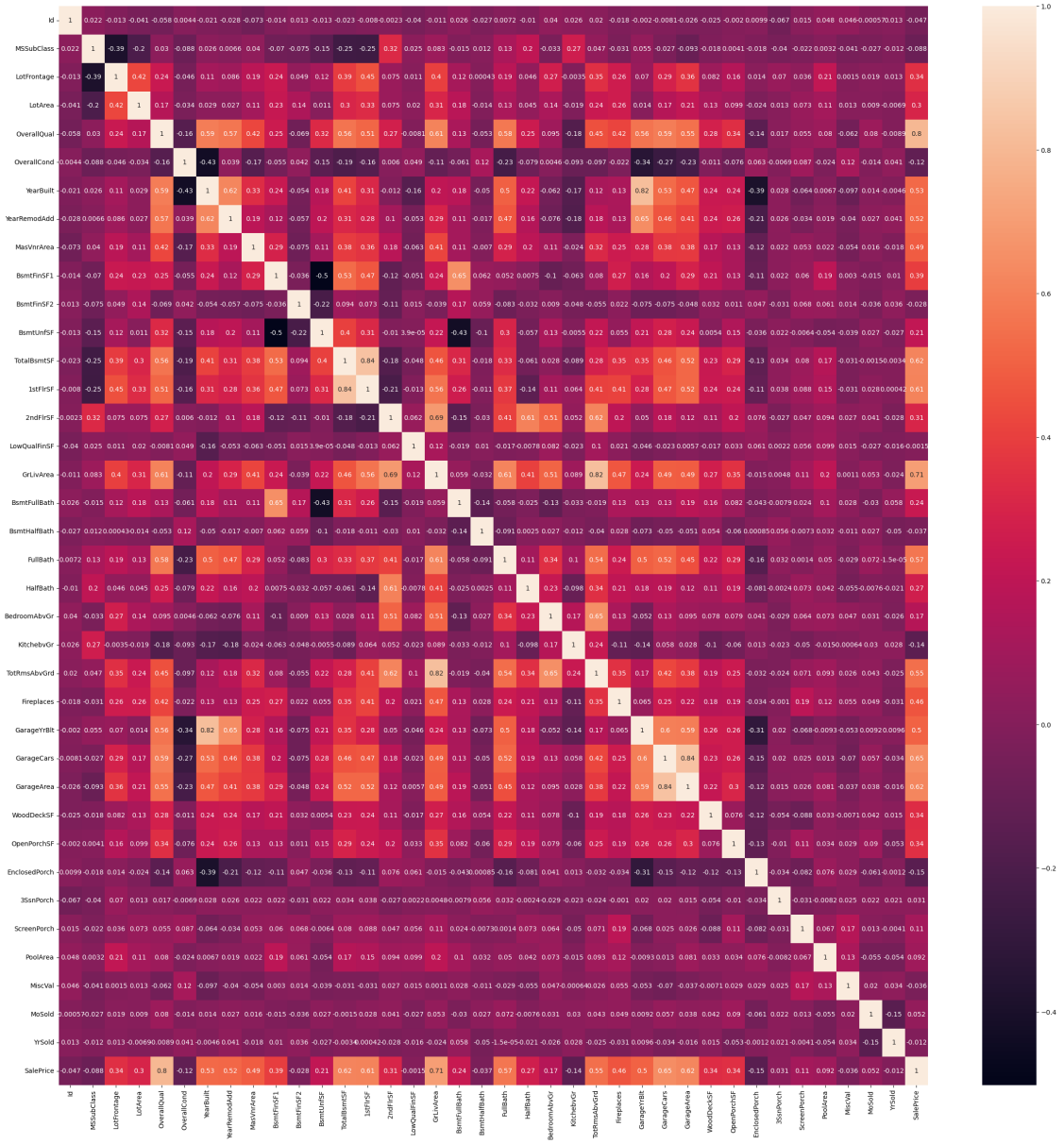
GarageYrBlt	…	0.255916	0.257141	-0.308278	0.019842
GarageCars	…	0.234276	0.258137	-0.151886	0.020141
GarageArea	…	0.223955	0.302558	-0.115749	0.015306
WoodDeckSF	…	1.000000	0.075525	-0.121061	-0.053825
OpenPorchSF	…	0.075525	1.000000	-0.130566	-0.010351
EnclosedPorch	…	-0.121061	-0.130566	1.000000	-0.034376
3SsnPorch	…	-0.053825	-0.010351	-0.034376	1.000000
ScreenPorch	…	-0.087575	0.112443	-0.081550	-0.031359
PoolArea	…	0.033076	0.033786	0.076342	-0.008215
MiscVal	…	-0.007101	0.028843	0.028795	0.024614
MoSold	…	0.041547	0.089767	-0.061083	0.022260
YrSold	…	0.014810	-0.053035	-0.001185	0.020731
SalePrice	…	0.336855	0.343354	-0.154843	0.030777

	ScreenPorch	PoolArea	MiscVal	MoSold	YrSold	SalePrice
Id	0.015183	0.048010	0.045799	-0.000570	0.013407	-0.047122
MSSubClass	-0.021789	0.003166	-0.040689	-0.027170	-0.012448	-0.088032
LotFrontage	0.035906	0.211746	0.001471	0.018815	0.013267	0.344270
LotArea	0.072517	0.109147	0.012790	0.008998	-0.006904	0.299962
OverallQual	0.055296	0.080131	-0.062064	0.079895	-0.008903	0.797881
OverallCond	0.087030	-0.023566	0.119772	-0.014236	0.041003	-0.124391
YearBuilt	-0.063694	0.006717	-0.096973	0.013784	-0.004585	0.525394
YearRemodAdd	-0.034288	0.019307	-0.040420	0.026884	0.041302	0.521253
MasVnrArea	0.052646	0.021648	-0.054044	0.015850	-0.017569	0.488658
BsmtFinSF1	0.059635	0.194349	0.003027	-0.015281	0.010224	0.390301
BsmtFinSF2	0.067899	0.061212	0.014290	-0.036101	0.036395	-0.028021
BsmtUnfSF	-0.006398	-0.053894	-0.038915	0.027068	-0.026736	0.213129
TotalBsmtSF	0.080259	0.171489	-0.031076	-0.001498	-0.003377	0.615612
1stFlrSF	0.087580	0.151761	-0.030909	0.027731	0.000420	0.607969
2ndFlrSF	0.047039	0.094076	0.027047	0.041485	-0.028010	0.306879
LowQualFinSF	0.056472	0.099089	0.015231	-0.026645	-0.016253	-0.001482
GrLivArea	0.108453	0.198551	0.001067	0.053071	-0.024436	0.705154
BsmtFullBath	0.023857	0.104349	0.027641	-0.030282	0.058467	0.236737
BsmtHalfBath	-0.007323	0.031503	-0.010947	0.027037	-0.049851	-0.036513
FullBath	0.001415	0.049608	-0.029397	0.072305	-0.000015	0.566627
HalfBath	0.073306	0.042099	-0.055379	-0.007637	-0.020875	0.268560
BedroomAbvGr	0.063660	0.073361	0.046523	0.031268	-0.026035	0.166814
KitchebvGr	-0.050308	-0.015114	-0.000640	0.030001	0.028463	-0.140497
TotRmsAbvGrd	0.070894	0.093387	0.026495	0.043097	-0.024812	0.547067
Fireplaces	0.192129	0.117108	0.054826	0.048788	-0.031402	0.461873
GarageYrBlt	-0.067596	-0.009295	-0.053295	0.009233	0.009596	0.504753
GarageCars	0.025135	0.012829	-0.069592	0.057481	-0.033507	0.647034
GarageArea	0.026446	0.080871	-0.036993	0.037597	-0.016206	0.619330
WoodDeckSF	-0.087575	0.033076	-0.007101	0.041547	0.014810	0.336855
OpenPorchSF	0.112443	0.033786	0.028843	0.089767	-0.053035	0.343354
EnclosedPorch	-0.081550	0.076342	0.028795	-0.061083	-0.001185	-0.154843
3SsnPorch	-0.031359	-0.008215	0.024614	0.022260	0.020731	0.030777

```
ScreenPorch       1.000000  0.067356  0.169857  0.012859 -0.004118   0.110427
PoolArea          0.067356  1.000000  0.128684 -0.054872 -0.053888   0.092488
MiscVal           0.169857  0.128684  1.000000  0.020067  0.034106  -0.036041
MoSold            0.012859 -0.054872  0.020067  1.000000 -0.150577   0.051568
YrSold           -0.004118 -0.053888  0.034106 -0.150577  1.000000  -0.011869
SalePrice         0.110427  0.092488 -0.036041  0.051568 -0.011869   1.000000

[38 rows x 38 columns]
```

[24]:
```
plt.figure(figsize=(30,30))
sns.heatmap(df_num.corr(),annot=True)
```

[24]: <AxesSubplot:>

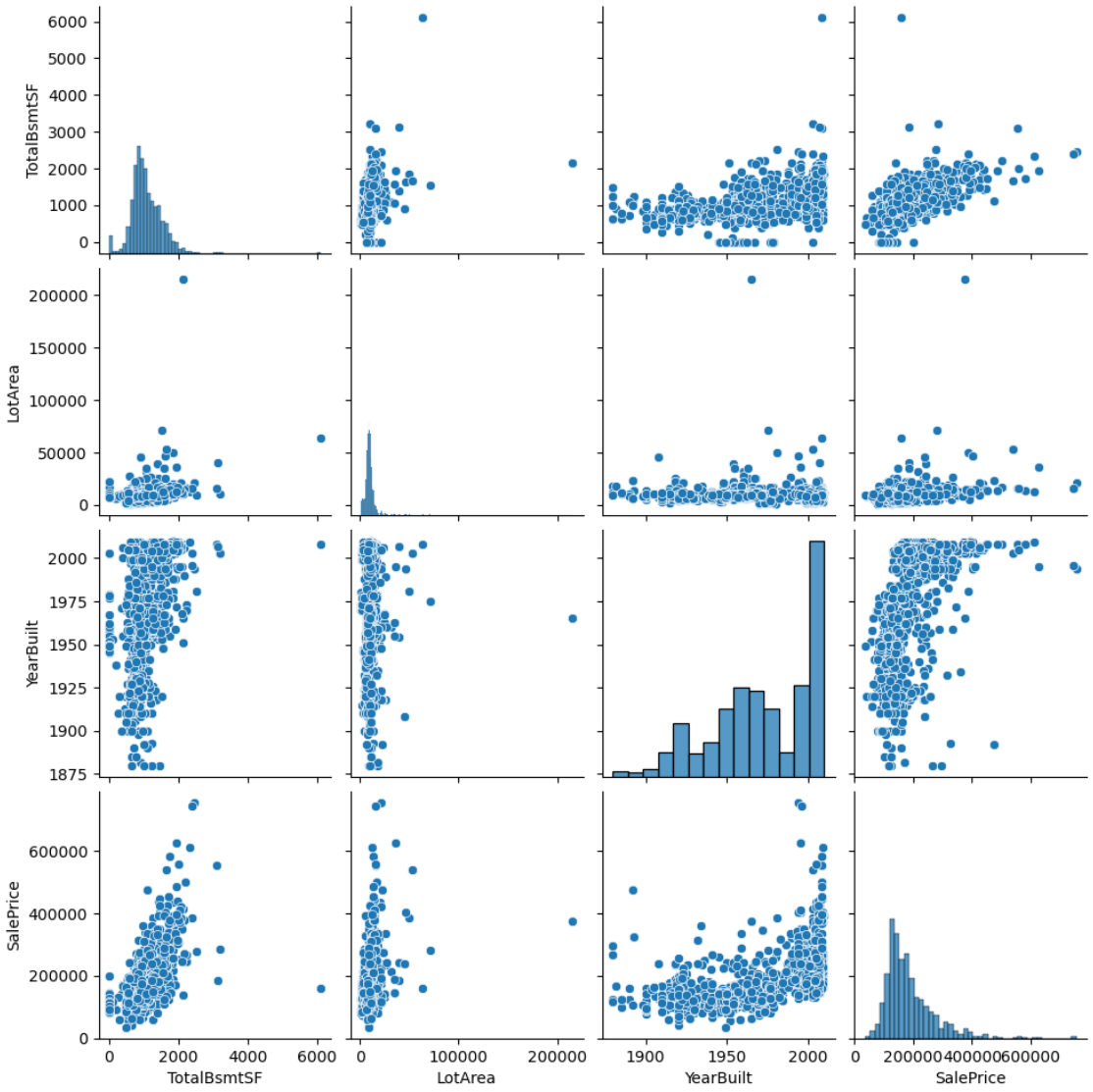


5 3. d. Pair plot for distribution and density

[25]:
```
plt.figure(figsize=(15,8))
cols=['TotalBsmtSF','LotArea','YearBuilt','SalePrice']
sns.pairplot(df_num,vars=cols)
```

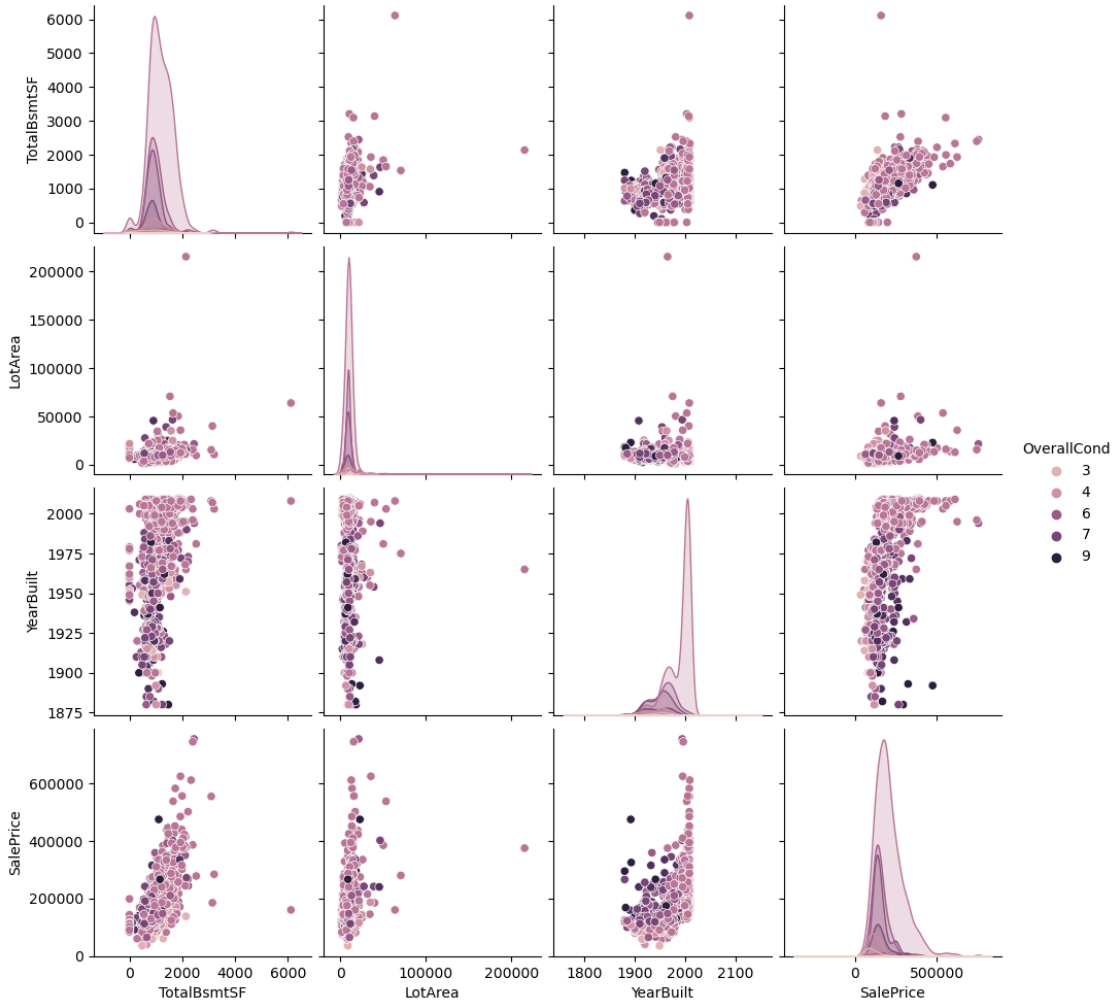
[25]: <seaborn.axisgrid.PairGrid at 0x2c8c3a9c2e0>

<Figure size 1500x800 with 0 Axes>



[26]:
```
plt.figure(figsize=(15,8))
cols=['TotalBsmtSF','LotArea','YearBuilt','SalePrice']
sns.pairplot(df_num,vars=cols,hue='OverallCond')
```

[26]:
```
<seaborn.axisgrid.PairGrid at 0x2c8c18b1340>

<Figure size 1500x800 with 0 Axes>
```



6 4. EDA of categorical variables

a. Missing value treatment

b. Count plot and box plot for bivariate analysis

c. Identify significant variables using p-values and Chi-Square values

[27]:
```
df_obj.head()
```

[27]:
```
  MSZoning Street Alley LotShape LandContour Utilities LotConfig LandSlope  \
0       RL   Pave   NaN      Reg         Lvl    AllPub    Inside       Gtl
1       RL   Pave   NaN      Reg         Lvl    AllPub       FR2       Gtl
2       RL   Pave   NaN      IR1         Lvl    AllPub    Inside       Gtl
3       RL   Pave   NaN      IR1         Lvl    AllPub    Corner       Gtl
4       RL   Pave   NaN      IR1         Lvl    AllPub       FR2       Gtl

  Neighborhood Condition1  … GarageType GarageFinish GarageQual GarageCond  \
0      CollgCr       Norm  …     Attchd          RFn         TA         TA
1      Veenker      Feedr  …     Attchd          RFn         TA         TA
2      CollgCr       Norm  …     Attchd          RFn         TA         TA
3      Crawfor       Norm  …     Detchd          Unf         TA         TA
4      NoRidge       Norm  …     Attchd          RFn         TA         TA

  PavedDrive PoolQC Fence MiscFeature SaleType SaleCondition
0          Y    NaN   NaN         NaN       WD        Normal
1          Y    NaN   NaN         NaN       WD        Normal
2          Y    NaN   NaN         NaN       WD        Normal
3          Y    NaN   NaN         NaN       WD       Abnorml
4          Y    NaN   NaN         NaN       WD        Normal

[5 rows x 43 columns]
```

6.1 4. a. Missing value treatment

[28]:
```
df_obj.isnull().sum()
```

[28]:
```
MSZoning           0
Street             0
Alley           1369
LotShape           0
LandContour        0
Utilities          0
LotConfig          0
LandSlope          0
Neighborhood       0
Condition1         0
Condition2         0
BldgType           0
HouseStyle         0
RoofStyle          0
RoofMatl           0
Exterior1st        0
Exterior2nd        0
MasVnrType         8
ExterQual          0
ExterCond          0
```

```
Foundation          0
BsmtQual           37
BsmtCond           37
BsmtExposure       38
BsmtFinType1       37
BsmtFinType2       38
Heating             0
HeatingQC           0
CentralAir          0
Electrical          1
KitchenQual         0
Functiol            0
FireplaceQu       690
GarageType         81
GarageFinish       81
GarageQual         81
GarageCond         81
PavedDrive          0
PoolQC           1453
Fence            1179
MiscFeature      1406
SaleType            0
SaleCondition       0
dtype: int64
```

[29]:
```
df_obj.
  ↪drop(['Alley','PoolQC','Fence','MiscFeature','FireplaceQu'],axis=1,inplace=True)
```

[30]:
```
df_obj.isnull().sum()
```

[30]:
```
MSZoning          0
Street            0
LotShape          0
LandContour       0
Utilities         0
LotConfig         0
LandSlope         0
Neighborhood      0
Condition1        0
Condition2        0
BldgType          0
HouseStyle        0
RoofStyle         0
RoofMatl          0
Exterior1st       0
Exterior2nd       0
MasVnrType        8
```

```
ExterQual          0
ExterCond          0
Foundation         0
BsmtQual          37
BsmtCond          37
BsmtExposure      38
BsmtFinType1      37
BsmtFinType2      38
Heating            0
HeatingQC          0
CentralAir         0
Electrical         1
KitchenQual        0
Functiol           0
GarageType        81
GarageFinish      81
GarageQual        81
GarageCond        81
PavedDrive         0
SaleType           0
SaleCondition      0
dtype: int64
```

[31]:
```
df_obj.dropna(inplace=True)
```

[32]:
```
df_obj.head()
```

[32]:

	MSZoning	Street	LotShape	LandContour	Utilities	LotConfig	LandSlope	\
0	RL	Pave	Reg	Lvl	AllPub	Inside	Gtl	
1	RL	Pave	Reg	Lvl	AllPub	FR2	Gtl	
2	RL	Pave	IR1	Lvl	AllPub	Inside	Gtl	
3	RL	Pave	IR1	Lvl	AllPub	Corner	Gtl	
4	RL	Pave	IR1	Lvl	AllPub	FR2	Gtl	

	Neighborhood	Condition1	Condition2	…	Electrical	KitchenQual	Functiol	\
0	CollgCr	Norm	Norm	…	SBrkr	Gd	Typ	
1	Veenker	Feedr	Norm	…	SBrkr	TA	Typ	
2	CollgCr	Norm	Norm	…	SBrkr	Gd	Typ	
3	Crawfor	Norm	Norm	…	SBrkr	Gd	Typ	
4	NoRidge	Norm	Norm	…	SBrkr	Gd	Typ	

	GarageType	GarageFinish	GarageQual	GarageCond	PavedDrive	SaleType	\
0	Attchd	RFn	TA	TA	Y	WD	
1	Attchd	RFn	TA	TA	Y	WD	
2	Attchd	RFn	TA	TA	Y	WD	
3	Detchd	Unf	TA	TA	Y	WD	
4	Attchd	RFn	TA	TA	Y	WD	

```
  SaleCondition
0        Normal
1        Normal
2        Normal
3       Abnorml
4        Normal

[5 rows x 38 columns]
```

[33]:
```
df_obj.isnull().sum()
```

[33]:
```
MSZoning         0
Street           0
LotShape         0
LandContour      0
Utilities        0
LotConfig        0
LandSlope        0
Neighborhood     0
Condition1       0
Condition2       0
BldgType         0
HouseStyle       0
RoofStyle        0
RoofMatl         0
Exterior1st      0
Exterior2nd      0
MasVnrType       0
ExterQual        0
ExterCond        0
Foundation       0
BsmtQual         0
BsmtCond         0
BsmtExposure     0
BsmtFinType1     0
BsmtFinType2     0
Heating          0
HeatingQC        0
CentralAir       0
Electrical       0
KitchenQual      0
Functiol         0
GarageType       0
GarageFinish     0
GarageQual       0
GarageCond       0
```

```
PavedDrive       0
SaleType         0
SaleCondition    0
dtype: int64
```

[34]:
```
df_obj.shape
```

[34]:
```
(1338, 38)
```

6.2 4. b. Count plot and box plot for bivariate analysis

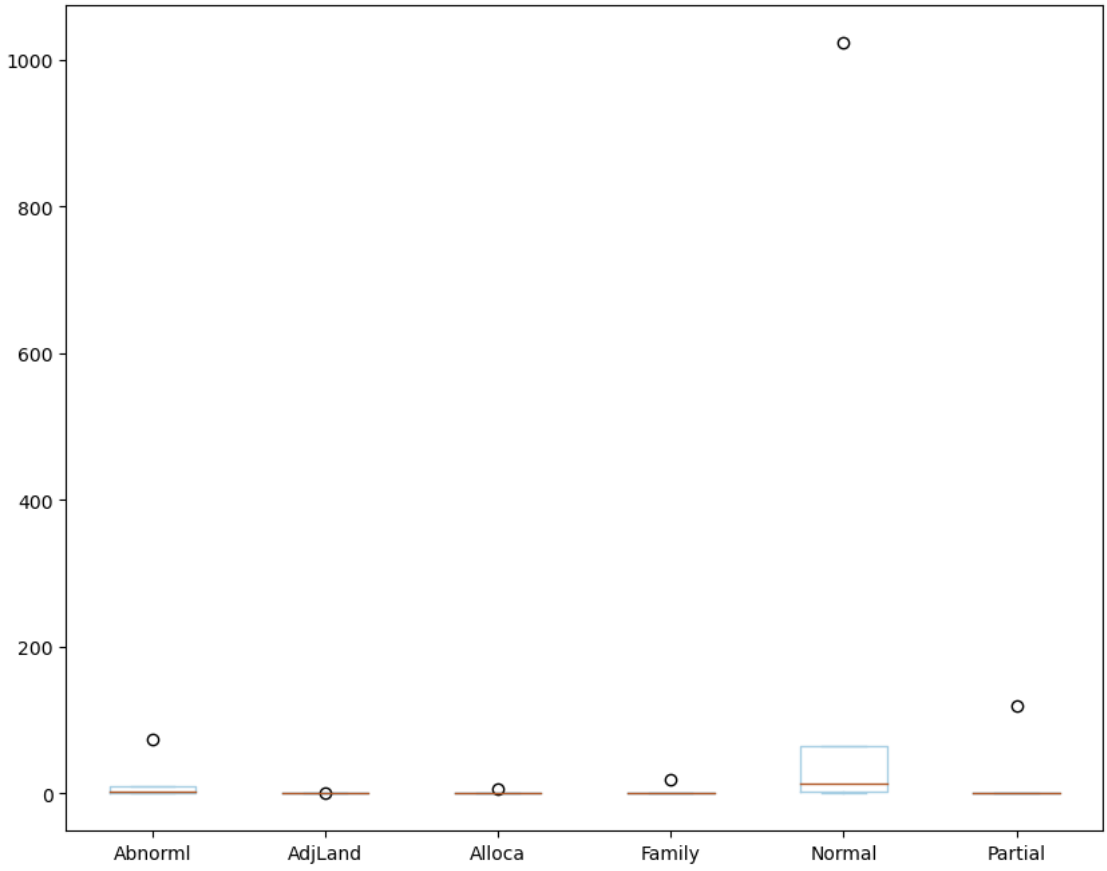
[35]:
```
df_obj.columns
```

[35]:
```
Index(['MSZoning', 'Street', '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',
       'Functiol', 'GarageType', 'GarageFinish', 'GarageQual', 'GarageCond',
       'PavedDrive', 'SaleType', 'SaleCondition'],
      dtype='object')
```

[36]:
```
crosstab=pd.crosstab(index=df_obj['Electrical'],columns=df_obj['SaleCondition'])
```

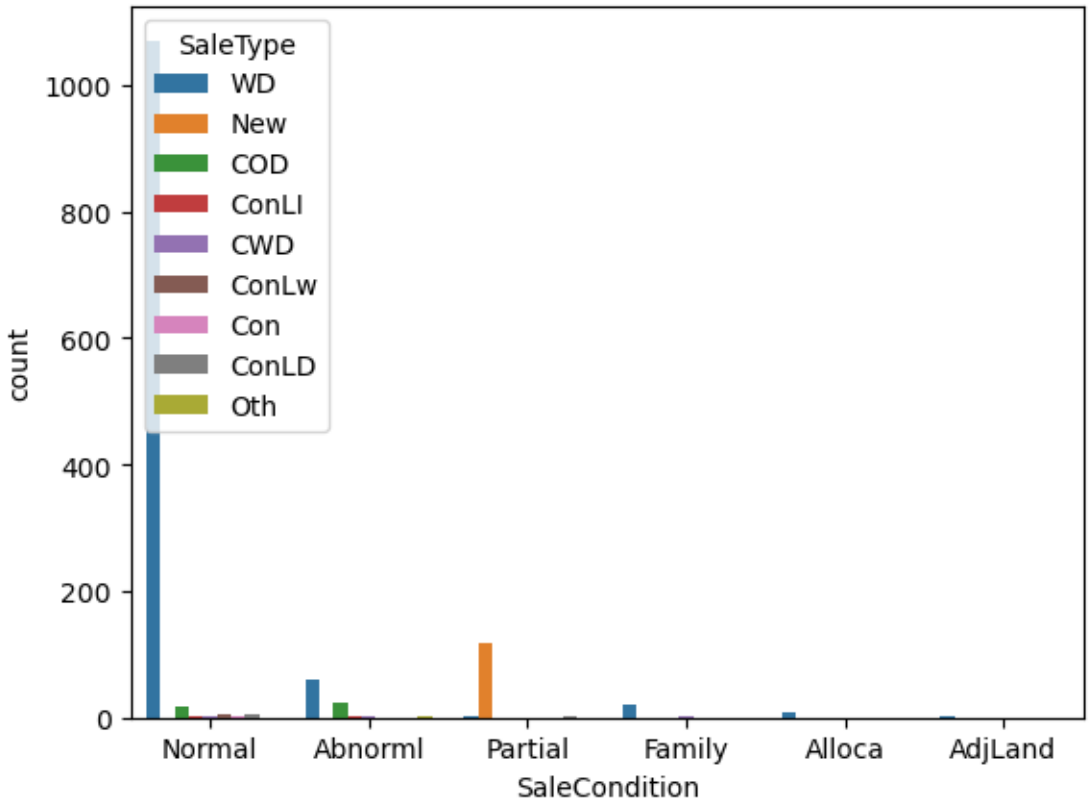
[37]:
```
crosstab.plot(kind='box',figsize=(10,8),stacked=True,colormap='Paired')
```

[37]:
```
<AxesSubplot:>
```



[38]:
```
sns.countplot(data=df_obj,x='SaleCondition',hue='SaleType')
```

[38]:
```
<AxesSubplot:xlabel='SaleCondition', ylabel='count'>
```



6.3 4. c. Identify significant variables using p-values and Chi-Square values

[40]:
```
import scipy.stats
from scipy.stats import chi2
```

[41]:
```
ct_table=pd.crosstab(df_obj['SaleCondition'],df_obj['SaleType'])
print('contingency_table :\n',ct_table)
```

```
contingency_table :
 SaleType        COD  CWD  Con  ConLD  ConLI  ConLw  New  Oth    WD
SaleCondition
Abnorml          24    1    0      0      1      0    0    1    59
AdjLand           0    0    0      0      0      0    0    0     1
Alloca            0    0    0      0      0      0    0    0     7
Family            0    1    0      0      0      0    0    0    19
Normal           18    2    2      5      3      4    0    0  1070
Partial           0    0    0      1      0      0  117    0     2
```

[44]:
```
chi2_stat, p, dof, expected = scipy.stats.chi2_contingency(ct_table)

print(f"chi2 statistic:     {chi2_stat}")
```

```
print(f"p-value:            {p}")
print(f"degrees of freedom: {dof}")
print("expected frequencies:\n",expected)
```

```
chi2 statistic:     1522.601445269729
p-value:            1.1136083863268581e-293
degrees of freedom: 40
expected frequencies:
 [[2.69955157e+00 2.57100149e-01 1.28550075e-01 3.85650224e-01
  2.57100149e-01 2.57100149e-01 7.52017937e+00 6.42750374e-02
  7.44304933e+01]
 [3.13901345e-02 2.98953662e-03 1.49476831e-03 4.48430493e-03
  2.98953662e-03 2.98953662e-03 8.74439462e-02 7.47384155e-04
  8.65470852e-01]
 [2.19730942e-01 2.09267564e-02 1.04633782e-02 3.13901345e-02
  2.09267564e-02 2.09267564e-02 6.12107623e-01 5.23168909e-03
  6.05829596e+00]
 [6.27802691e-01 5.97907324e-02 2.98953662e-02 8.96860987e-02
  5.97907324e-02 5.97907324e-02 1.74887892e+00 1.49476831e-02
  1.73094170e+01]
 [3.46547085e+01 3.30044843e+00 1.65022422e+00 4.95067265e+00
  3.30044843e+00 3.30044843e+00 9.65381166e+01 8.25112108e-01
  9.55479821e+02]
 [3.76681614e+00 3.58744395e-01 1.79372197e-01 5.38116592e-01
  3.58744395e-01 3.58744395e-01 1.04932735e+01 8.96860987e-02
  1.03856502e+02]]
```

7 5. Combine all the significant categorical and numerical variables

[66]:
```
df_obj.columns
```

[66]:
```
Index(['MSZoning', 'Street', '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',
       'Functiol', 'GarageType', 'GarageFinish', 'GarageQual', 'GarageCond',
       'PavedDrive', 'SaleType', 'SaleCondition'],
      dtype='object')
```

[62]:
```
from sklearn.preprocessing import LabelEncoder
le=LabelEncoder()
```

[68]:
```
OneHot=df_obj['Condition1'].str.get_dummies('|')
```

[69]:
```
OneHot
```

[69]:

	Artery	Feedr	Norm	PosA	PosN	RRAe	RRAn	RRNe	RRNn
0	0	0	1	0	0	0	0	0	0
1	0	1	0	0	0	0	0	0	0
2	0	0	1	0	0	0	0	0	0
3	0	0	1	0	0	0	0	0	0
4	0	0	1	0	0	0	0	0	0
…	…	…	…	…	…	…	…	…	…
1455	0	0	1	0	0	0	0	0	0
1456	0	0	1	0	0	0	0	0	0
1457	0	0	1	0	0	0	0	0	0
1458	0	0	1	0	0	0	0	0	0
1459	0	0	1	0	0	0	0	0	0

[1338 rows x 9 columns]

[70]:
```
df_com=df_obj.copy()
```

[71]:
```
df_com=pd.concat([df_obj,OneHot],axis=1)
```

[73]:
```
df_com
```

[73]:

	MSZoning	Street	LotShape	LandContour	Utilities	LotConfig	LandSlope	\
0	RL	Pave	Reg	Lvl	AllPub	Inside	Gtl	
1	RL	Pave	Reg	Lvl	AllPub	FR2	Gtl	
2	RL	Pave	IR1	Lvl	AllPub	Inside	Gtl	
3	RL	Pave	IR1	Lvl	AllPub	Corner	Gtl	
4	RL	Pave	IR1	Lvl	AllPub	FR2	Gtl	
…	…	…	…	…	…	…	…	…
1455	RL	Pave	Reg	Lvl	AllPub	Inside	Gtl	
1456	RL	Pave	Reg	Lvl	AllPub	Inside	Gtl	
1457	RL	Pave	Reg	Lvl	AllPub	Inside	Gtl	
1458	RL	Pave	Reg	Lvl	AllPub	Inside	Gtl	
1459	RL	Pave	Reg	Lvl	AllPub	Inside	Gtl	

	Neighborhood	Condition1	Condition2	…	SaleCondition	Artery	Feedr	Norm	\
0	CollgCr	Norm	Norm	…	Normal	0	0	1	
1	Veenker	Feedr	Norm	…	Normal	0	1	0	
2	CollgCr	Norm	Norm	…	Normal	0	0	1	
3	Crawfor	Norm	Norm	…	Abnorml	0	0	1	
4	NoRidge	Norm	Norm	…	Normal	0	0	1	
…	…	…	…	…	…	…	…	…	…
1455	Gilbert	Norm	Norm	…	Normal	0	0	1	
1456	NWAmes	Norm	Norm	…	Normal	0	0	1	
1457	Crawfor	Norm	Norm	…	Normal	0	0	1	
1458	mes	Norm	Norm	…	Normal	0	0	1	

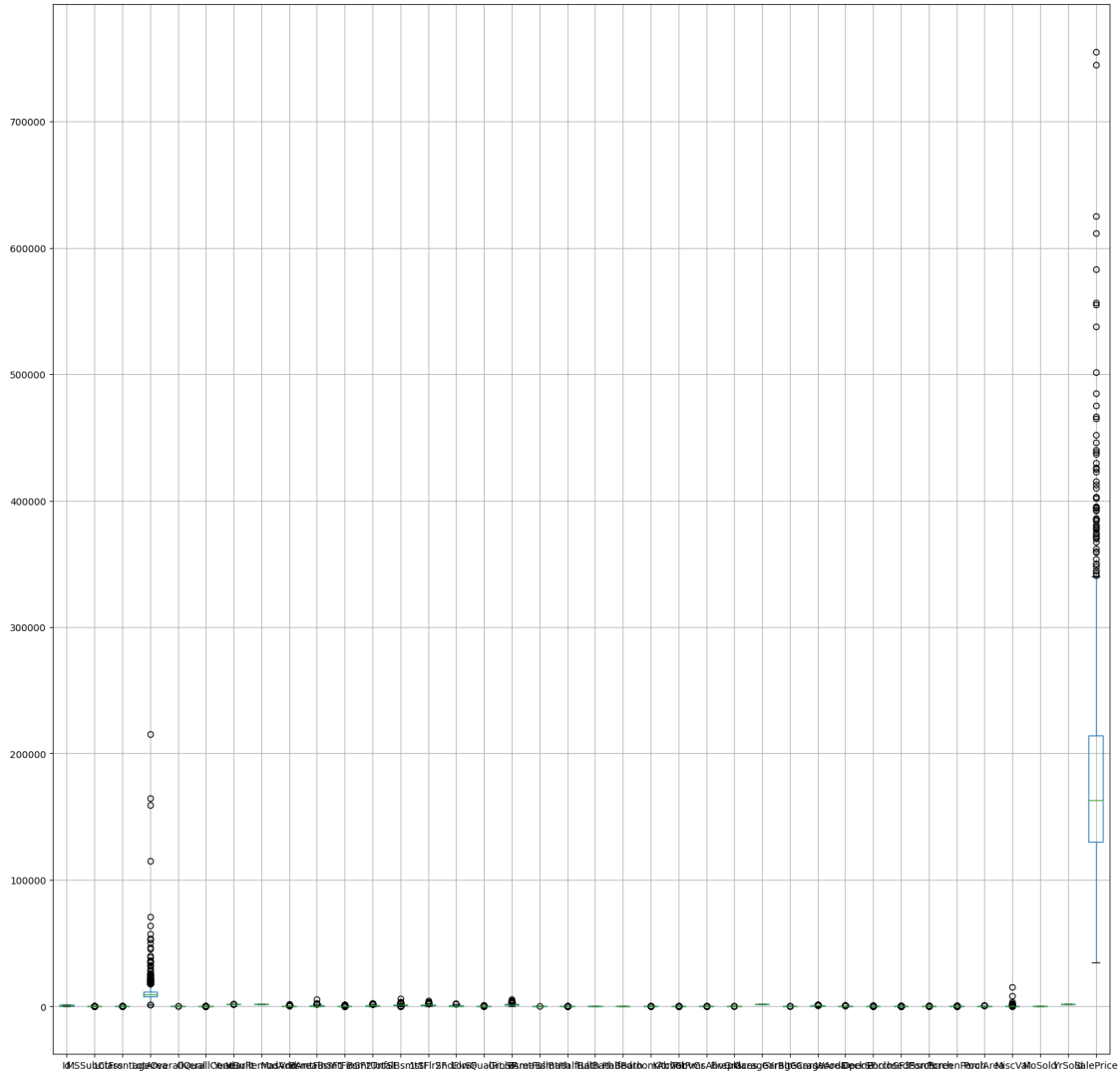
```
1459      Edwards       Norm       Norm  …        Normal      0     0    1

     PosA PosN RRAe RRAn RRNe RRNn
0       0    0    0    0    0    0
1       0    0    0    0    0    0
2       0    0    0    0    0    0
3       0    0    0    0    0    0
4       0    0    0    0    0    0
…    …   …   …   …   …   …   …
1455    0    0    0    0    0    0
1456    0    0    0    0    0    0
1457    0    0    0    0    0    0
1458    0    0    0    0    0    0
1459    0    0    0    0    0    0

[1338 rows x 47 columns]
```

8 6. Plot box plot for the new dataset to find the variables with outliers

[75]:
```
df_outlier=df.copy()
```

[77]:
```
plt.figure(figsize=(20,20))
df_outlier.boxplot()
```

[77]:
```
<AxesSubplot:>
```


700000
600000
500000
400000
300000
200000
100000
0
SalePrice

[ ]: