Exercise

For this exercise, you will be working with the House Price Dataset.

Please grab the train.csv file from Kaggle and explore this dataset. You need to perform explroatory data analysis and see if there is any correlation between the variables and analyze the distribution of the dataset. The question is open-ended and basically you're asked to perform EDA.

1- Write a summary of your findings in one page (e.g., summary statistics, plots) and submit the pdf file. Therefore, for part 3 of your assignment, you need to submit at least one jupyter notebook file and one pdf file.

2- Push your code and project to github and provide the link to your code here. Ensure that your github project is organized to at least couple of main folders, ensure that you have the README file as well:

- Src
- Data
- Docs
- Results

Read this link for further info:
https://gist.github.com/ericmjl/27e50331f24db3e8f957d1fe7bbbe510

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

train_df = pd.read_csv("train.csv")

train_df.head()
```

Out[1]:

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

5 rows × 81 columns

In [15]:
```
train_df.isnull().sum()
```

Out[15]:
```
MSSubClass         0
MSZoning           0
LotFrontage      259
LotArea            0
Street             0
                ...
MoSold             0
YrSold             0
SaleType           0
SaleCondition      0
SalePrice          0
Length: 80, dtype: int64
```

In [16]:
```
train_df.dtypes
```

Out[16]:
```
MSSubClass         int64
MSZoning          object
LotFrontage      float64
LotArea            int64
Street            object
                  ...
MoSold             int64
YrSold             int64
SaleType          object
SaleCondition     object
SalePrice          int64
Length: 80, dtype: object
```

In [ ]:
```
#drop the Id column as it is not necessary for analysis
```

In [4]:
```
train_df = train_df.drop('Id', axis=1)
train_df.head()
```

Out[4]:

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

5 rows × 80 columns

In [5]:
```
train_df.info() # check features of dataset
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1460 entries, 0 to 1459
Data columns (total 80 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   MSSubClass     1460 non-null   int64
 1   MSZoning       1460 non-null   object
 2   LotFrontage    1201 non-null   float64
 3   LotArea        1460 non-null   int64
 4   Street         1460 non-null   object
 5   Alley          91 non-null     object
 6   LotShape       1460 non-null   object
 7   LandContour    1460 non-null   object
 8   Utilities      1460 non-null   object
 9   LotConfig      1460 non-null   object
 10  LandSlope      1460 non-null   object
 11  Neighborhood   1460 non-null   object
 12  Condition1     1460 non-null   object
 13  Condition2     1460 non-null   object
 14  BldgType       1460 non-null   object
 15  HouseStyle     1460 non-null   object
 16  OverallQual    1460 non-null   int64
 17  OverallCond    1460 non-null   int64
 18  YearBuilt      1460 non-null   int64
 19  YearRemodAdd   1460 non-null   int64
 20  RoofStyle      1460 non-null   object
 21  RoofMatl       1460 non-null   object
 22  Exterior1st    1460 non-null   object
 23  Exterior2nd    1460 non-null   object
 24  MasVnrType     1452 non-null   object
 25  MasVnrArea     1452 non-null   float64
 26  ExterQual      1460 non-null   object
 27  ExterCond      1460 non-null   object
 28  Foundation     1460 non-null   object
 29  BsmtQual       1423 non-null   object
 30  BsmtCond       1423 non-null   object
 31  BsmtExposure   1422 non-null   object
 32  BsmtFinType1   1423 non-null   object
 33  BsmtFinSF1     1460 non-null   int64
 34  BsmtFinType2   1422 non-null   object
 35  BsmtFinSF2     1460 non-null   int64
 36  BsmtUnfSF      1460 non-null   int64
 37  TotalBsmtSF    1460 non-null   int64
 38  Heating        1460 non-null   object
 39  HeatingQC      1460 non-null   object
 40  CentralAir     1460 non-null   object
 41  Electrical     1459 non-null   object
 42  1stFlrSF       1460 non-null   int64
 43  2ndFlrSF       1460 non-null   int64
 44  LowQualFinSF   1460 non-null   int64
 45  GrLivArea      1460 non-null   int64
 46  BsmtFullBath   1460 non-null   int64
 47  BsmtHalfBath   1460 non-null   int64
 48  FullBath       1460 non-null   int64
 49  HalfBath       1460 non-null   int64
 50  BedroomAbvGr   1460 non-null   int64
 51  KitchenAbvGr   1460 non-null   int64
 52  KitchenQual    1460 non-null   object
 53  TotRmsAbvGrd   1460 non-null   int64
 54  Functional     1460 non-null   object
```

```
 55  Fireplaces     1460 non-null   int64
 56  FireplaceQu    770 non-null    object
 57  GarageType     1379 non-null   object
 58  GarageYrBlt    1379 non-null   float64
 59  GarageFinish   1379 non-null   object
 60  GarageCars     1460 non-null   int64
 61  GarageArea     1460 non-null   int64
 62  GarageQual     1379 non-null   object
 63  GarageCond     1379 non-null   object
 64  PavedDrive     1460 non-null   object
 65  WoodDeckSF     1460 non-null   int64
 66  OpenPorchSF    1460 non-null   int64
 67  EnclosedPorch  1460 non-null   int64
 68  3SsnPorch      1460 non-null   int64
 69  ScreenPorch    1460 non-null   int64
 70  PoolArea       1460 non-null   int64
 71  PoolQC         7 non-null      object
 72  Fence          281 non-null    object
 73  MiscFeature    54 non-null     object
 74  MiscVal        1460 non-null   int64
 75  MoSold         1460 non-null   int64
 76  YrSold         1460 non-null   int64
 77  SaleType       1460 non-null   object
 78  SaleCondition  1460 non-null   object
 79  SalePrice      1460 non-null   int64
dtypes: float64(3), int64(34), object(43)
memory usage: 912.6+ KB
```

In [14]:
```
train_df.describe() # summary of data
```

Out[14]:

	MSSubClass	LotFrontage	LotArea	OverallQual	OverallCond	YearBuilt	YearRemodAd
count	1460.000000	1201.000000	1460.000000	1460.000000	1460.000000	1460.000000	1460.0000
mean	56.897260	70.049958	10516.828082	6.099315	5.575342	1971.267808	1984.8657
std	42.300571	24.284752	9981.264932	1.382997	1.112799	30.202904	20.6454
min	20.000000	21.000000	1300.000000	1.000000	1.000000	1872.000000	1950.0000
25%	20.000000	59.000000	7553.500000	5.000000	5.000000	1954.000000	1967.0000
50%	50.000000	69.000000	9478.500000	6.000000	5.000000	1973.000000	1994.0000
75%	70.000000	80.000000	11601.500000	7.000000	6.000000	2000.000000	2004.0000
max	190.000000	313.000000	215245.000000	10.000000	9.000000	2010.000000	2010.0000

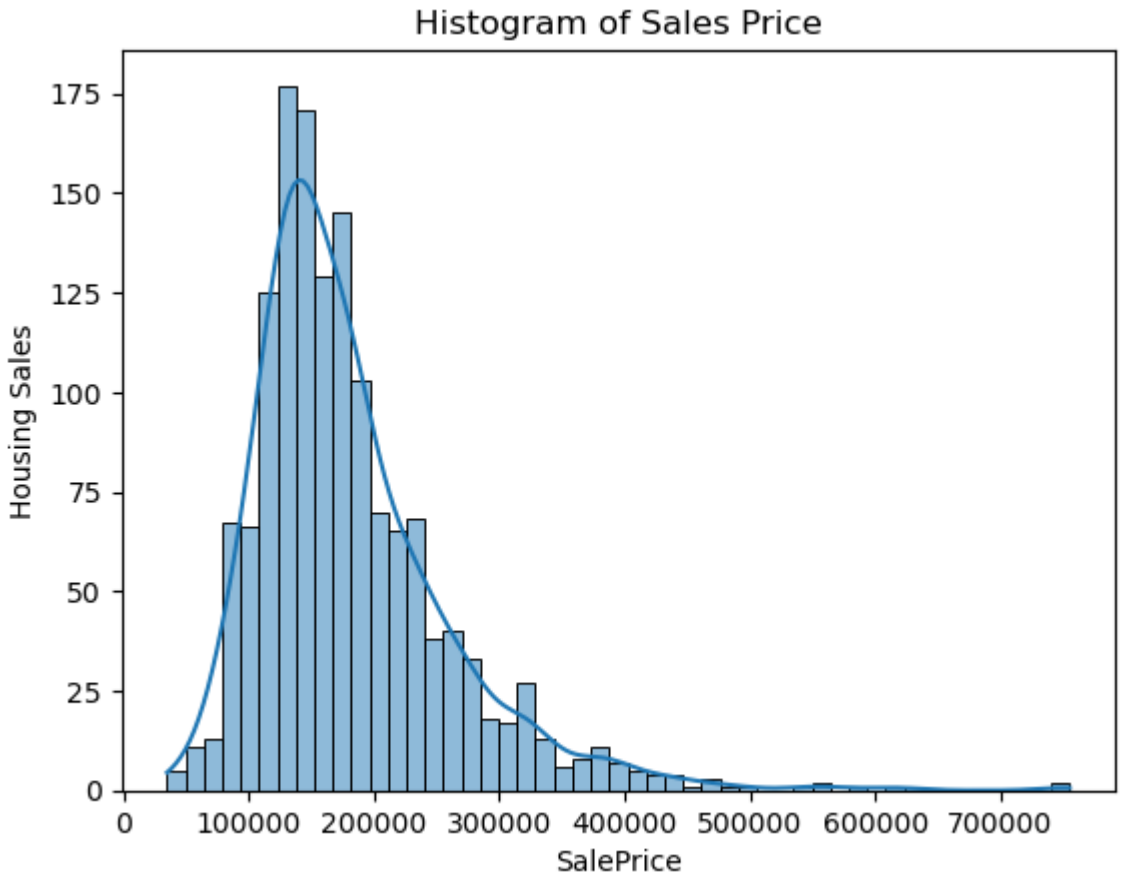
8 rows × 37 columns

In [ ]:
```
# House Price Distribution
```

In [6]:
```
print(train_df['SalePrice'].describe())
```

```
count      1460.000000
mean     180921.195890
std       79442.502883
min       34900.000000
25%      129975.000000
50%      163000.000000
75%      214000.000000
max      755000.000000
Name: SalePrice, dtype: float64
```

In [39]:
```
sale_price = train_df['SalePrice']

sns.histplot(sale_price, kde=True)

plt.ylabel("Housing Sales")
plt.title("Histogram of Sales Price")

plt.show()
sale_price.describe()
```



Out[39]:
```
count      1460.000000
mean     180921.195890
std       79442.502883
min       34900.000000
25%      129975.000000
50%      163000.000000
75%      214000.000000
max      755000.000000
Name: SalePrice, dtype: float64
```

In [47]:
```
### positive skewed(right)
### Max number of houses sold lies between $100k to $200k
```

In [ ]:
```
# Numerical data distribution
```

In [9]:
```
list(set(train_df.dtypes.tolist()))
```

Out[9]:
```
[dtype('int64'), dtype('float64'), dtype('O')]
```

In [10]:
```
df_num = train_df.select_dtypes(include = ['float64', 'int64'])
df_num.head()
```

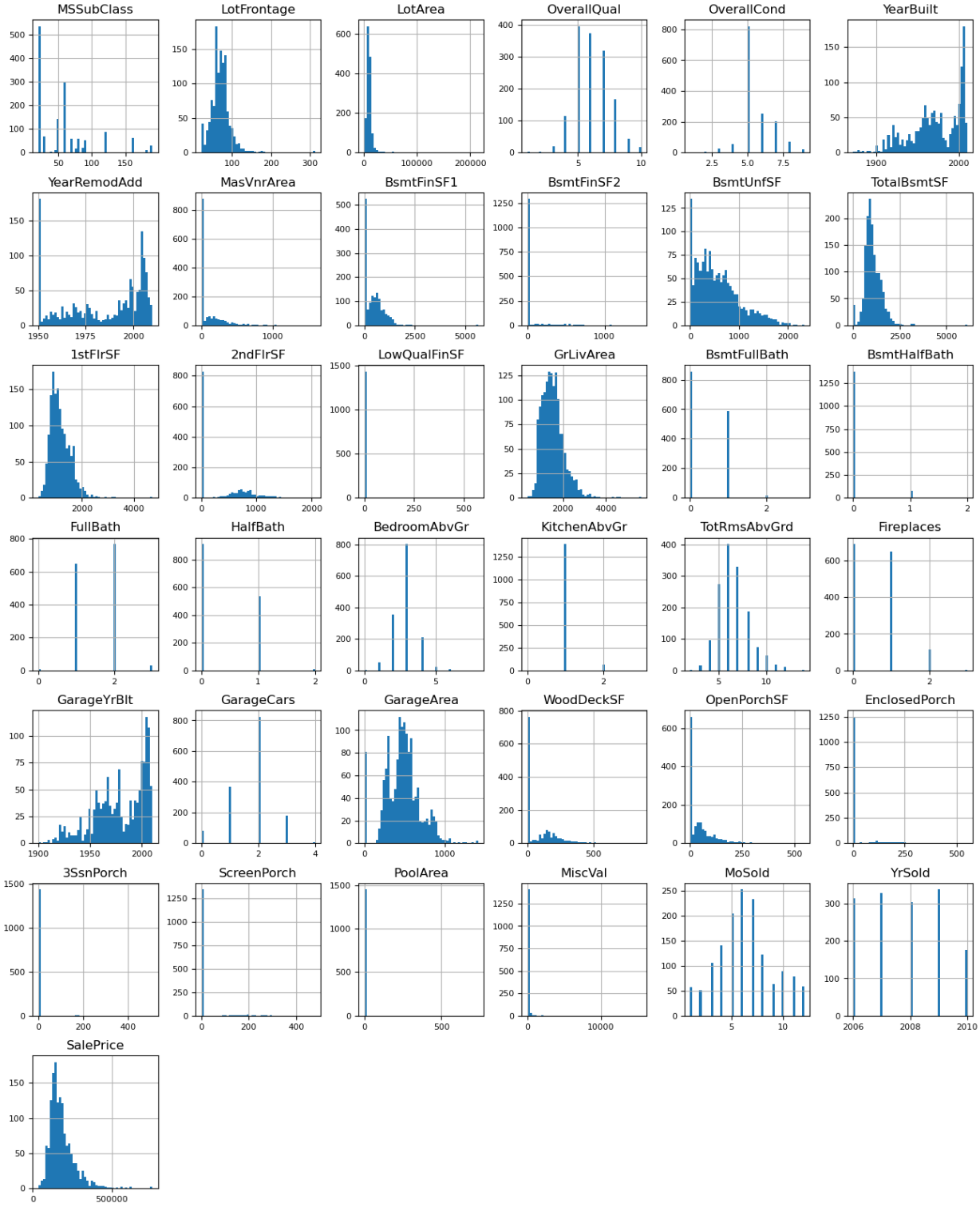
Out[10]:

	MSSubClass	LotFrontage	LotArea	OverallQual	OverallCond	YearBuilt	YearRemodAdd	MasVnrAr
0	60	65.0	8450	7	5	2003	2003	196
1	20	80.0	9600	6	8	1976	1976	0
2	60	68.0	11250	7	5	2001	2002	162
3	70	60.0	9550	7	5	1915	1970	0
4	60	84.0	14260	8	5	2000	2000	350

5 rows × 37 columns

In [ ]:
```
# plot the distribution for all the numerical features
```

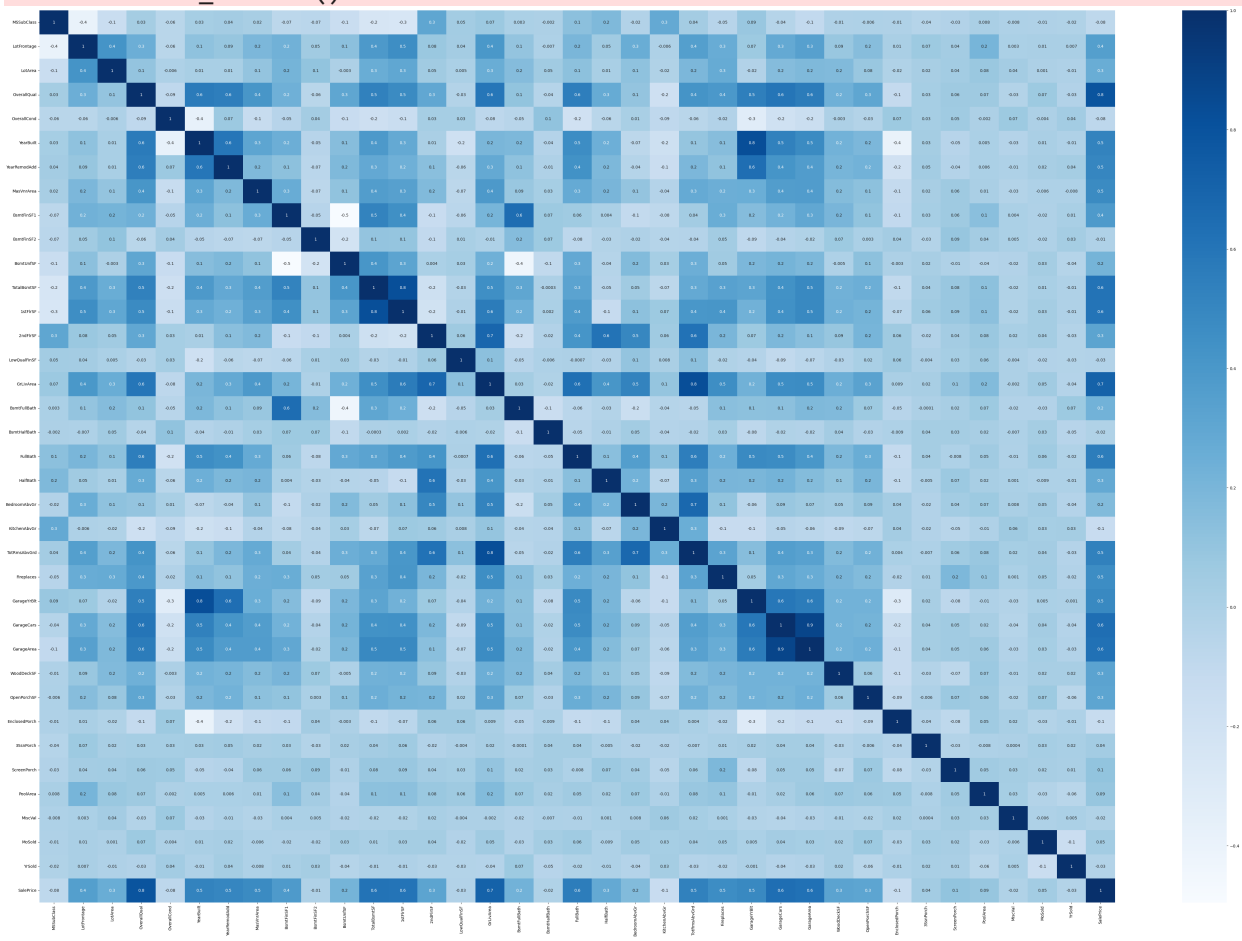
In [11]:
```
df_num.hist(figsize=(16, 20), bins=50, xlabelsize=8, ylabelsize=8);
```



In [ ]:
```
# Heatmap to see correlation
```

In [18]:
```
cor = train_df.corr()
plt.figure(figsize=(60,40))
sns.heatmap(data=cor,annot=True,cmap='Blues',fmt='.1g')
plt.show()
```

```
C:\Users\Sahib\AppData\Local\Temp\ipykernel_20536\1514070793.py:1: FutureWarning: The default value of numeric_only in DataFrame.corr is deprecated. In a future version, it will default to False. Select only valid columns or specify the value of numeric_only to silence this warning.
  cor = train_df.corr()
```



In [ ]:
```
# Correlation Matrix
```

In [19]:
```
print(cor)
```

	MSSubClass	LotFrontage	LotArea	OverallQual	OverallCond	\
MSSubClass	1.000000	-0.386347	-0.139781	0.032628	-0.059316	
LotFrontage	-0.386347	1.000000	0.426095	0.251646	-0.059213	
LotArea	-0.139781	0.426095	1.000000	0.105806	-0.005636	
OverallQual	0.032628	0.251646	0.105806	1.000000	-0.091932	
OverallCond	-0.059316	-0.059213	-0.005636	-0.091932	1.000000	
YearBuilt	0.027850	0.123349	0.014228	0.572323	-0.375983	
YearRemodAdd	0.040581	0.088866	0.013788	0.550684	0.073741	
MasVnrArea	0.022936	0.193458	0.104160	0.411876	-0.128101	
BsmtFinSF1	-0.069836	0.233633	0.214103	0.239666	-0.046231	
BsmtFinSF2	-0.065649	0.049900	0.111170	-0.059119	0.040229	
BsmtUnfSF	-0.140759	0.132644	-0.002618	0.308159	-0.136841	
TotalBsmtSF	-0.238518	0.392075	0.260833	0.537808	-0.171098	
1stFlrSF	-0.251758	0.457181	0.299475	0.476224	-0.144203	
2ndFlrSF	0.307886	0.080177	0.050986	0.295493	0.028942	
LowQualFinSF	0.046474	0.038469	0.004779	-0.030429	0.025494	
GrLivArea	0.074853	0.402797	0.263116	0.593007	-0.079686	
BsmtFullBath	0.003491	0.100949	0.158155	0.111098	-0.054942	
BsmtHalfBath	-0.002333	-0.007234	0.048046	-0.040150	0.117821	
FullBath	0.131608	0.198769	0.126031	0.550600	-0.194149	
HalfBath	0.177354	0.053532	0.014259	0.273458	-0.060769	
BedroomAbvGr	-0.023438	0.263170	0.119690	0.101676	0.012980	
KitchenAbvGr	0.281721	-0.006069	-0.017784	-0.183882	-0.087001	
TotRmsAbvGrd	0.040380	0.352096	0.190015	0.427452	-0.057583	
Fireplaces	-0.045569	0.266639	0.271364	0.396765	-0.023820	
GarageYrBlt	0.085072	0.070250	-0.024947	0.547766	-0.324297	
GarageCars	-0.040110	0.285691	0.154871	0.600671	-0.185758	
GarageArea	-0.098672	0.344997	0.180403	0.562022	-0.151521	
WoodDeckSF	-0.012579	0.088521	0.171698	0.238923	-0.003334	
OpenPorchSF	-0.006100	0.151972	0.084774	0.308819	-0.032589	
EnclosedPorch	-0.012037	0.010700	-0.018340	-0.113937	0.070356	
3SsnPorch	-0.043825	0.070029	0.020423	0.030371	0.025504	
ScreenPorch	-0.026030	0.041383	0.043160	0.064886	0.054811	
PoolArea	0.008283	0.206167	0.077672	0.065166	-0.001985	
MiscVal	-0.007683	0.003368	0.038068	-0.031406	0.068777	
MoSold	-0.013585	0.011200	0.001205	0.070815	-0.003511	
YrSold	-0.021407	0.007450	-0.014261	-0.027347	0.043950	
SalePrice	-0.084284	0.351799	0.263843	0.790982	-0.077856	

	YearBuilt	YearRemodAdd	MasVnrArea	BsmtFinSF1	BsmtFinSF2	\
MSSubClass	0.027850	0.040581	0.022936	-0.069836	-0.065649	
LotFrontage	0.123349	0.088866	0.193458	0.233633	0.049900	
LotArea	0.014228	0.013788	0.104160	0.214103	0.111170	
OverallQual	0.572323	0.550684	0.411876	0.239666	-0.059119	
OverallCond	-0.375983	0.073741	-0.128101	-0.046231	0.040229	
YearBuilt	1.000000	0.592855	0.315707	0.249503	-0.049107	
YearRemodAdd	0.592855	1.000000	0.179618	0.128451	-0.067759	
MasVnrArea	0.315707	0.179618	1.000000	0.264736	-0.072319	
BsmtFinSF1	0.249503	0.128451	0.264736	1.000000	-0.050117	
BsmtFinSF2	-0.049107	-0.067759	-0.072319	-0.050117	1.000000	
BsmtUnfSF	0.149040	0.181133	0.114442	-0.495251	-0.209294	
TotalBsmtSF	0.391452	0.291066	0.363936	0.522396	0.104810	
1stFlrSF	0.281986	0.240379	0.344501	0.445863	0.097117	
2ndFlrSF	0.010308	0.140024	0.174561	-0.137079	-0.099260	
LowQualFinSF	-0.183784	-0.062419	-0.069071	-0.064503	0.014807	
GrLivArea	0.199010	0.287389	0.390857	0.208171	-0.009640	
BsmtFullBath	0.187599	0.119470	0.085310	0.649212	0.158678	
BsmtHalfBath	-0.038162	-0.012337	0.026673	0.067418	0.070948	
FullBath	0.468271	0.439046	0.276833	0.058543	-0.076444	
HalfBath	0.242656	0.183331	0.201444	0.004262	-0.032148	

BedroomAbvGr	-0.070651	-0.040581	0.102821	-0.107355	-0.015728
KitchenAbvGr	-0.174800	-0.149598	-0.037610	-0.081007	-0.040751
TotRmsAbvGrd	0.095589	0.191740	0.280682	0.044316	-0.035227
Fireplaces	0.147716	0.112581	0.249070	0.260011	0.046921
GarageYrBlt	0.825667	0.642277	0.252691	0.153484	-0.088011
GarageCars	0.537850	0.420622	0.364204	0.224054	-0.038264
GarageArea	0.478954	0.371600	0.373066	0.296970	-0.018227
WoodDeckSF	0.224880	0.205726	0.159718	0.204306	0.067898
OpenPorchSF	0.188686	0.226298	0.125703	0.111761	0.003093
EnclosedPorch	-0.387268	-0.193919	-0.110204	-0.102303	0.036543
3SsnPorch	0.031355	0.045286	0.018796	0.026451	-0.029993
ScreenPorch	-0.050364	-0.038740	0.061466	0.062021	0.088871
PoolArea	0.004950	0.005829	0.011723	0.140491	0.041709
MiscVal	-0.034383	-0.010286	-0.029815	0.003571	0.004940
MoSold	0.012398	0.021490	-0.005965	-0.015727	-0.015211
YrSold	-0.013618	0.035743	-0.008201	0.014359	0.031706
SalePrice	0.522897	0.507101	0.477493	0.386420	-0.011378

	...	WoodDeckSF	OpenPorchSF	EnclosedPorch	3SsnPorch	\
MSSubClass	...	-0.012579	-0.006100	-0.012037	-0.043825	
LotFrontage	...	0.088521	0.151972	0.010700	0.070029	
LotArea	...	0.171698	0.084774	-0.018340	0.020423	
OverallQual	...	0.238923	0.308819	-0.113937	0.030371	
OverallCond	...	-0.003334	-0.032589	0.070356	0.025504	
YearBuilt	...	0.224880	0.188686	-0.387268	0.031355	
YearRemodAdd	...	0.205726	0.226298	-0.193919	0.045286	
MasVnrArea	...	0.159718	0.125703	-0.110204	0.018796	
BsmtFinSF1	...	0.204306	0.111761	-0.102303	0.026451	
BsmtFinSF2	...	0.067898	0.003093	0.036543	-0.029993	
BsmtUnfSF	...	-0.005316	0.129005	-0.002538	0.020764	
TotalBsmtSF	...	0.232019	0.247264	-0.095478	0.037384	
1stFlrSF	...	0.235459	0.211671	-0.065292	0.056104	
2ndFlrSF	...	0.092165	0.208026	0.061989	-0.024358	
LowQualFinSF	...	-0.025444	0.018251	0.061081	-0.004296	
GrLivArea	...	0.247433	0.330224	0.009113	0.020643	
BsmtFullBath	...	0.175315	0.067341	-0.049911	-0.000106	
BsmtHalfBath	...	0.040161	-0.025324	-0.008555	0.035114	
FullBath	...	0.187703	0.259977	-0.115093	0.035353	
HalfBath	...	0.108080	0.199740	-0.095317	-0.004972	
BedroomAbvGr	...	0.046854	0.093810	0.041570	-0.024478	
KitchenAbvGr	...	-0.090130	-0.070091	0.037312	-0.024600	
TotRmsAbvGrd	...	0.165984	0.234192	0.004151	-0.006683	
Fireplaces	...	0.200019	0.169405	-0.024822	0.011257	
GarageYrBlt	...	0.224577	0.228425	-0.297003	0.023544	
GarageCars	...	0.226342	0.213569	-0.151434	0.035765	
GarageArea	...	0.224666	0.241435	-0.121777	0.035087	
WoodDeckSF	...	1.000000	0.058661	-0.125989	-0.032771	
OpenPorchSF	...	0.058661	1.000000	-0.093079	-0.005842	
EnclosedPorch	...	-0.125989	-0.093079	1.000000	-0.037305	
3SsnPorch	...	-0.032771	-0.005842	-0.037305	1.000000	
ScreenPorch	...	-0.074181	0.074304	-0.082864	-0.031436	
PoolArea	...	0.073378	0.060762	0.054203	-0.007992	
MiscVal	...	-0.009551	-0.018584	0.018361	0.000354	
MoSold	...	0.021011	0.071255	-0.028887	0.029474	
YrSold	...	0.022270	-0.057619	-0.009916	0.018645	
SalePrice	...	0.324413	0.315856	-0.128578	0.044584	

	ScreenPorch	PoolArea	MiscVal	MoSold	YrSold	SalePrice
MSSubClass	-0.026030	0.008283	-0.007683	-0.013585	-0.021407	-0.084284
LotFrontage	0.041383	0.206167	0.003368	0.011200	0.007450	0.351799

```
LotArea           0.043160  0.077672  0.038068  0.001205 -0.014261   0.263843
OverallQual       0.064886  0.065166 -0.031406  0.070815 -0.027347   0.790982
OverallCond       0.054811 -0.001985  0.068777 -0.003511  0.043950  -0.077856
YearBuilt        -0.050364  0.004950 -0.034383  0.012398 -0.013618   0.522897
YearRemodAdd     -0.038740  0.005829 -0.010286  0.021490  0.035743   0.507101
MasVnrArea        0.061466  0.011723 -0.029815 -0.005965 -0.008201   0.477493
BsmtFinSF1        0.062021  0.140491  0.003571 -0.015727  0.014359   0.386420
BsmtFinSF2        0.088871  0.041709  0.004940 -0.015211  0.031706  -0.011378
BsmtUnfSF        -0.012579 -0.035092 -0.023837  0.034888 -0.041258   0.214479
TotalBsmtSF       0.084489  0.126053 -0.018479  0.013196 -0.014969   0.613581
1stFlrSF          0.088758  0.131525 -0.021096  0.031372 -0.013604   0.605852
2ndFlrSF          0.040606  0.081487  0.016197  0.035164 -0.028700   0.319334
LowQualFinSF      0.026799  0.062157 -0.003793 -0.022174 -0.028921  -0.025606
GrLivArea         0.101510  0.170205 -0.002416  0.050240 -0.036526   0.708624
BsmtFullBath      0.023148  0.067616 -0.023047 -0.025361  0.067049   0.227122
BsmtHalfBath      0.032121  0.020025 -0.007367  0.032873 -0.046524  -0.016844
FullBath         -0.008106  0.049604 -0.014290  0.055872 -0.019669   0.560664
HalfBath          0.072426  0.022381  0.001290 -0.009050 -0.010269   0.284108
BedroomAbvGr      0.044300  0.070703  0.007767  0.046544 -0.036014   0.168213
KitchenAbvGr     -0.051613 -0.014525  0.062341  0.026589  0.031687  -0.135907
TotRmsAbvGrd      0.059383  0.083757  0.024763  0.036907 -0.034516   0.533723
Fireplaces        0.184530  0.095074  0.001409  0.046357 -0.024096   0.466929
GarageYrBlt      -0.075418 -0.014501 -0.032417  0.005337 -0.001014   0.486362
GarageCars        0.050494  0.020934 -0.043080  0.040522 -0.039117   0.640409
GarageArea        0.051412  0.061047 -0.027400  0.027974 -0.027378   0.623431
WoodDeckSF       -0.074181  0.073378 -0.009551  0.021011  0.022270   0.324413
OpenPorchSF       0.074304  0.060762 -0.018584  0.071255 -0.057619   0.315856
EnclosedPorch    -0.082864  0.054203  0.018361 -0.028887 -0.009916  -0.128578
3SsnPorch        -0.031436 -0.007992  0.000354  0.029474  0.018645   0.044584
ScreenPorch       1.000000  0.051307  0.031946  0.023217  0.010694   0.111447
PoolArea          0.051307  1.000000  0.029669 -0.033737 -0.059689   0.092404
MiscVal           0.031946  0.029669  1.000000 -0.006495  0.004906  -0.021190
MoSold            0.023217 -0.033737 -0.006495  1.000000 -0.145721   0.046432
YrSold            0.010694 -0.059689  0.004906 -0.145721  1.000000  -0.028923
SalePrice         0.111447  0.092404 -0.021190  0.046432 -0.028923   1.000000

[37 rows x 37 columns]
```

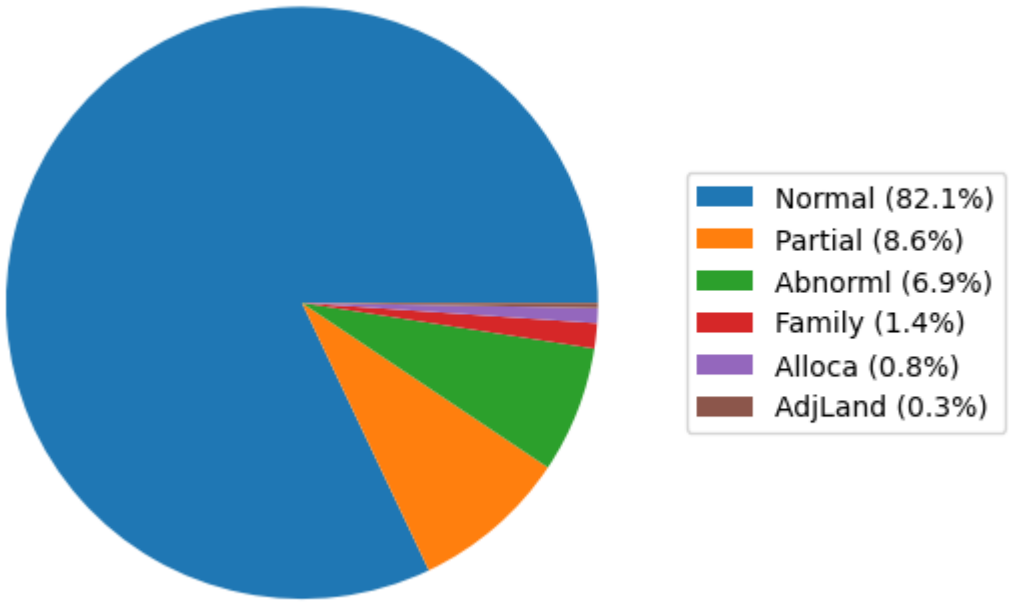
In [21]:
```
# Scatter plot of Gr living area and sale price
```

In [22]:
```
sns.lmplot(train_df, x="GrLivArea", y="SalePrice")

plt.xlabel("Living Area")
plt.ylabel("Sale Price")
plt.title("Scatter Plot of Living Area to Sales Price")

plt.show()
```

In [41]:
```
### There is a strong positive relation between Sales Privce and Living Area, espeacia
```

In [23]:
```
# Distribution of sale condition - Pie chart
```

In [24]:
```
sale_cond_rel_freq = train_df['SaleCondition'].value_counts(normalize=True) * 100

sale_cond_labels = [f'{label} ({percentage:.1f}%)' for label, percentage in
                    zip(sale_cond_rel_freq.index.values, sale_cond_rel_freq)]

plt.pie(sale_cond_rel_freq, labels=None)
plt.legend(labels=sale_cond_labels, loc='center left', bbox_to_anchor=(1, 0.5))

plt.show()
```



In [42]:
```
### Mostly sold house condtion was normal follwed by partial and abnormal. Normal is m
```

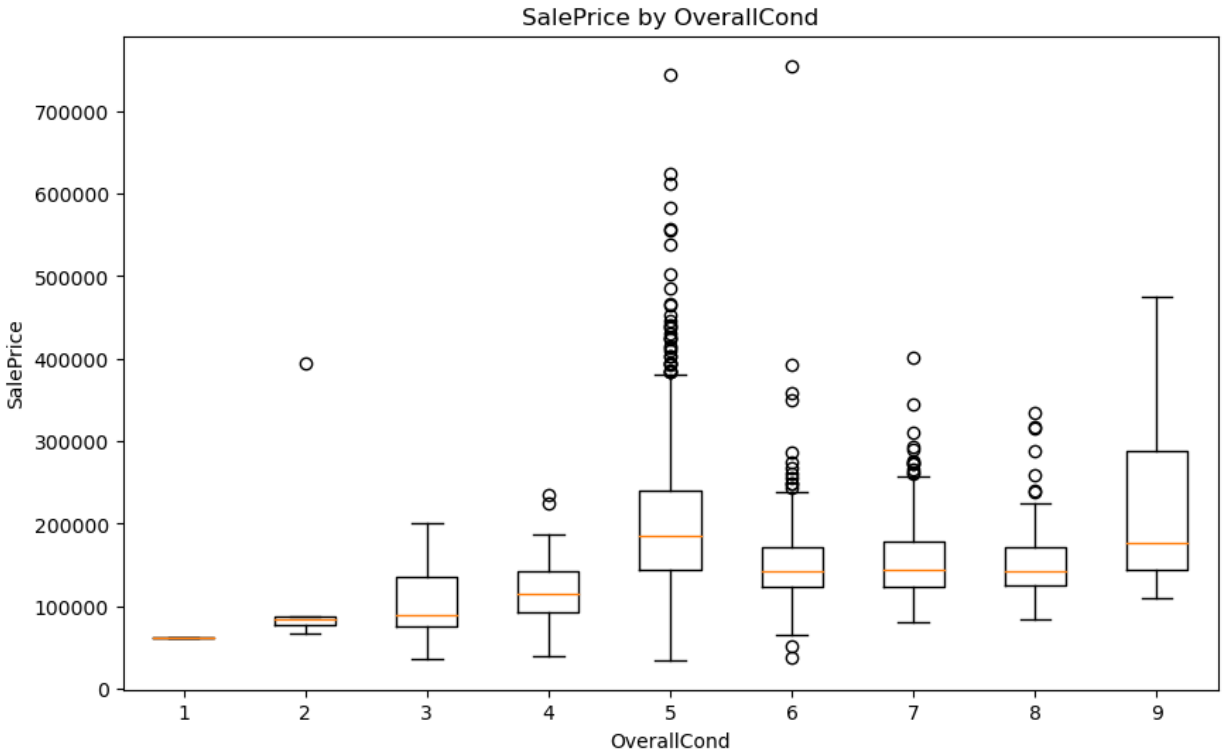
In [25]:
```
# House Proportion with and without pool
```

In [27]:
```
pool = train_df['PoolQC'].value_counts()

pool
```

Out[27]:
```
Gd    3
Ex    2
Fa    2
Name: PoolQC, dtype: int64
```

In [ ]:
```
# Sales Price by overall condition
```

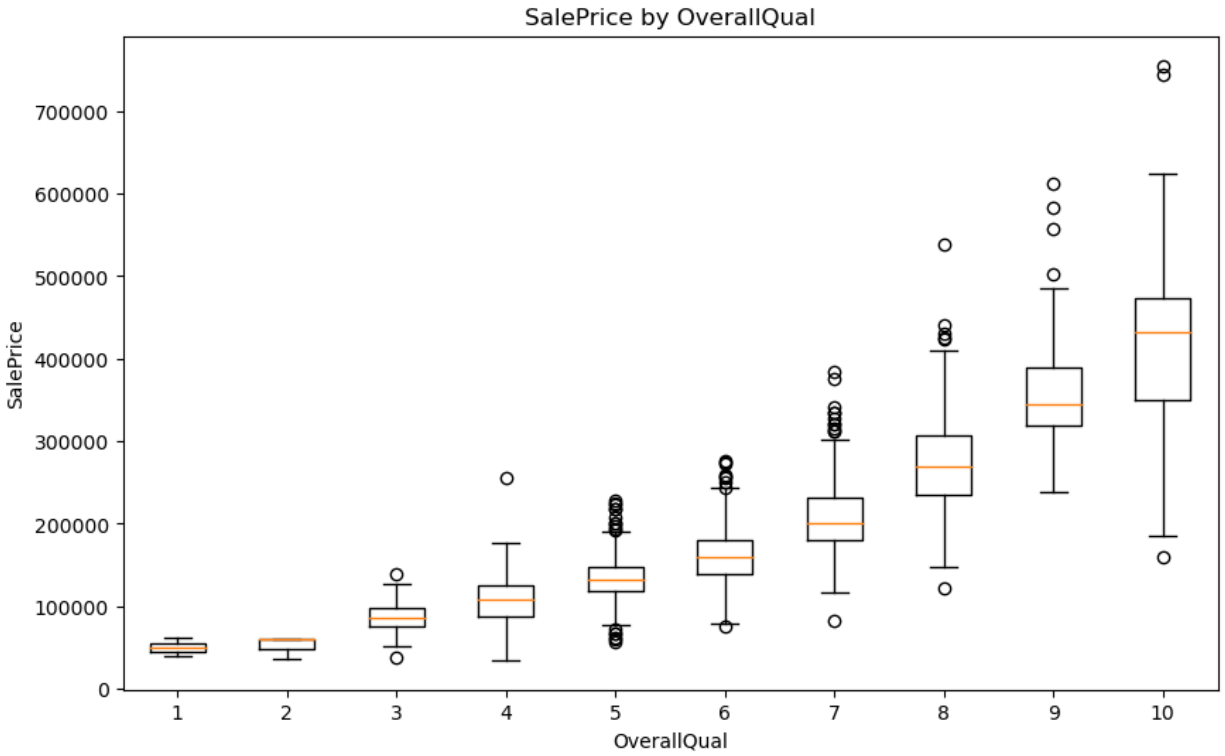
In [31]:
```
grouped_data = train_df.groupby('OverallCond')

# Create a list of data frames, one for each group
data_frames = [grouped_data.get_group(group) for group in grouped_data.groups]

# Create side-by-side box plots
plt.figure(figsize=(10, 6))
plt.boxplot([df['SalePrice'] for df in data_frames], labels=grouped_data.groups)
plt.xlabel('OverallCond')
plt.ylabel('SalePrice')
plt.title('SalePrice by OverallCond')
plt.show()
```



In [44]:
```
### For overall condition 5 and 9, sales prices are more influential and ranging betwe
### overall condition has most outliers
```

In [32]:
```
# Sales Price by overall Quality
```

In [33]:
```
grouped_data = train_df.groupby('OverallQual')

# Create a list of data frames, one for each group
data_frames = [grouped_data.get_group(group) for group in grouped_data.groups]

# Create side-by-side box plots
plt.figure(figsize=(10, 6))
plt.boxplot([df['SalePrice'] for df in data_frames], labels=grouped_data.groups)
plt.xlabel('OverallQual')
plt.ylabel('SalePrice')
plt.title('SalePrice by OverallQual')
plt.show()
```



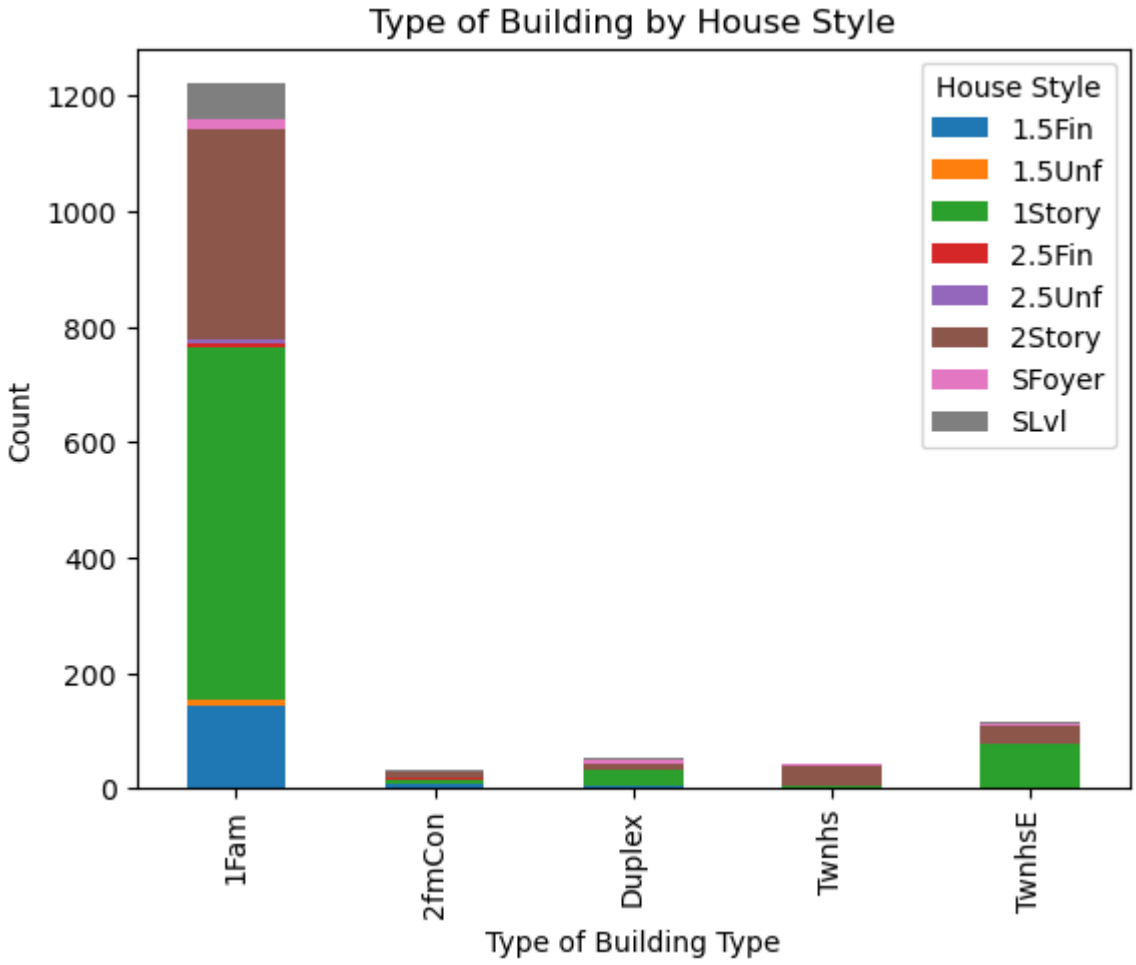
In [45]:
```
### Linear increasing relationship of sales prioce with overall quality. 10 overall qu
```

In [35]:
```
from scipy.stats import chi2_contingency

contingency_table = pd.crosstab(train_df['BldgType'], train_df['HouseStyle'])

# Perform chi-square test
chi2, p, dof, expected = chi2_contingency(contingency_table)

# Output the test results
print("Chi-square statistic:", chi2)
print("P-value:", p)
print("Degrees of freedom:", dof)
print("Expected frequencies table:")
print(expected)
```

```
Chi-square statistic: 170.37966418024408
P-value: 2.366736289460599e-22
Degrees of freedom: 28
Expected frequencies table:
[[1.28684932e+02 1.16986301e+01 6.06657534e+02 6.68493151e+00
  9.19178082e+00 3.71849315e+02 3.09178082e+01 5.43150685e+01]
 [3.26986301e+00 2.97260274e-01 1.54150685e+01 1.69863014e-01
  2.33561644e-01 9.44863014e+00 7.85616438e-01 1.38013699e+00]
 [5.48493151e+00 4.98630137e-01 2.58575342e+01 2.84931507e-01
  3.91780822e-01 1.58493151e+01 1.31780822e+00 2.31506849e+00]
 [4.53561644e+00 4.12328767e-01 2.13821918e+01 2.35616438e-01
  3.23972603e-01 1.31061644e+01 1.08972603e+00 1.91438356e+00]
 [1.20246575e+01 1.09315068e+00 5.66876712e+01 6.24657534e-01
  8.58904110e-01 3.47465753e+01 2.88904110e+00 5.07534247e+00]]
```

In [37]:
```
contingency_table.plot(kind='bar', stacked=True)
plt.xlabel('Type of Building Type')
plt.ylabel('Count')
plt.title('Type of Building by House Style')
```

```
plt.legend(title='House Style')
plt.show()
```



In [46]:
```
### 1 Family has more preference for differnt building house styles in which 1 story i
```

Final Summary

In summary, the analysis of the provided data yields the following key insights:

Skewed Distribution: The distribution of sale prices is positively skewed to the right, suggesting that the majority of houses sold fall in the lower price range.

Price Range: The most common price range for houses sold is between 100k and 200k

Positive Relation with Living Area: There is a strong positive relationship between sales price and living area, particularly for living areas between 1,000 to 3,000 square feet.

House Condition: The most frequently sold houses are in "normal" condition, followed by "partial" and "abnormal." Among these, "normal" condition houses have the most significant influence on sale prices.

Overall Condition Influence: Houses with overall conditions rated 5 and 9 tend to have more influence on sale prices, with prices ranging from 100,000 to 500,000.

Outliers: The overall condition variable has the most outliers, suggesting that there are exceptional cases where the condition significantly affects the sale price.

Quality and Price Relationship: There is a linearly increasing relationship between sale price and overall quality, with a broader spread of prices for houses with an overall quality rating of 10.

Dwelling Preferences: One-family dwellings tend to have a diverse preference for various building house styles. Among these styles, "1 story" homes are the most preferred, followed by "2 story" homes.