```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
import os
from scipy import stats
from scipy.stats import norm
from sklearn.preprocessing import StandardScaler
from sklearn.preprocessing import MinMaxScaler
from sklearn.preprocessing import LabelEncoder
from sklearn.preprocessing import PolynomialFeatures
from sklearn.impute import SimpleImputer
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.linear_model import LogisticRegression
import sklearn.metrics as metrics
from sklearn import preprocessing
from numpy import array
from sklearn.model_selection import KFold, cross_val_score
from sklearn.linear_model import Lasso
from sklearn.metrics import mean_squared_error
from sklearn.tree import DecisionTreeRegressor

%matplotlib inline
sns.set()
```

```
df_train = pd.read_csv("train.csv")
df_test = pd.read_csv("test.csv")
```

```
print(df_train.shape)
print("*"*50)
print(df_test.shape)
```

```
(1460, 81)
**************************************************
(1459, 80)
```

```
df_train.head()
```

```
   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
```

```
  LandContour Utilities  ... PoolArea PoolQC Fence MiscFeature MiscVal
MoSold  \
0         Lvl    AllPub  ...        0    NaN   NaN         NaN       0
2
1         Lvl    AllPub  ...        0    NaN   NaN         NaN       0
5
2         Lvl    AllPub  ...        0    NaN   NaN         NaN       0
9
3         Lvl    AllPub  ...        0    NaN   NaN         NaN       0
2
4         Lvl    AllPub  ...        0    NaN   NaN         NaN       0
12

  YrSold  SaleType  SaleCondition  SalePrice
0   2008        WD         Normal     208500
1   2007        WD         Normal     181500
2   2008        WD         Normal     223500
3   2006        WD        Abnorml     140000
4   2008        WD         Normal     250000

[5 rows x 81 columns]
```

```
df_test.head()
```

```
     Id  MSSubClass MSZoning  LotFrontage  LotArea Street Alley
LotShape  \
0  1461          20       RH         80.0    11622   Pave   NaN
Reg
1  1462          20       RL         81.0    14267   Pave   NaN
IR1
2  1463          60       RL         74.0    13830   Pave   NaN
IR1
3  1464          60       RL         78.0     9978   Pave   NaN
IR1
4  1465         120       RL         43.0     5005   Pave   NaN
IR1

  LandContour Utilities  ... ScreenPorch PoolArea PoolQC  Fence
MiscFeature  \
0         Lvl    AllPub  ...         120        0    NaN  MnPrv
NaN
1         Lvl    AllPub  ...           0        0    NaN    NaN
Gar2
2         Lvl    AllPub  ...           0        0    NaN  MnPrv
NaN
3         Lvl    AllPub  ...           0        0    NaN    NaN
NaN
4         HLS    AllPub  ...         144        0    NaN    NaN
NaN
```

```
  MiscVal MoSold  YrSold  SaleType  SaleCondition
0       0      6    2010        WD         Normal
1   12500      6    2010        WD         Normal
2       0      3    2010        WD         Normal
3       0      6    2010        WD         Normal
4       0      1    2010        WD         Normal

[5 rows x 80 columns]
```

EDA

```
df_train.describe()
```

```
                Id   MSSubClass  LotFrontage        LotArea
OverallQual  \
count  1460.000000  1460.000000  1201.000000    1460.000000
1460.000000
mean    730.500000    56.897260    70.049958   10516.828082
6.099315
std     421.610009    42.300571    24.284752    9981.264932
1.382997
min       1.000000    20.000000    21.000000    1300.000000
1.000000
25%     365.750000    20.000000    59.000000    7553.500000
5.000000
50%     730.500000    50.000000    69.000000    9478.500000
6.000000
75%    1095.250000    70.000000    80.000000   11601.500000
7.000000
max    1460.000000   190.000000   313.000000  215245.000000
10.000000

       OverallCond    YearBuilt  YearRemodAdd   MasVnrArea
BsmtFinSF1  ...  \
count  1460.000000  1460.000000   1460.000000  1452.000000
1460.000000  ...
mean      5.575342  1971.267808   1984.865753   103.685262
443.639726  ...
std       1.112799    30.202904     20.645407   181.066207
456.098091  ...
min       1.000000  1872.000000   1950.000000     0.000000
0.000000  ...
25%       5.000000  1954.000000   1967.000000     0.000000
0.000000  ...
50%       5.000000  1973.000000   1994.000000     0.000000
383.500000  ...
75%       6.000000  2000.000000   2004.000000   166.000000
712.250000  ...
max       9.000000  2010.000000   2010.000000  1600.000000
```

```
5644.000000  ...

        WoodDeckSF  OpenPorchSF  EnclosedPorch    3SsnPorch
ScreenPorch  \
count  1460.000000  1460.000000    1460.000000  1460.000000
1460.000000
mean     94.244521    46.660274      21.954110     3.409589
15.060959
std     125.338794    66.256028      61.119149    29.317331
55.757415
min       0.000000     0.000000       0.000000     0.000000
0.000000
25%       0.000000     0.000000       0.000000     0.000000
0.000000
50%       0.000000    25.000000       0.000000     0.000000
0.000000
75%     168.000000    68.000000       0.000000     0.000000
0.000000
max     857.000000   547.000000     552.000000   508.000000
480.000000

          PoolArea       MiscVal       MoSold       YrSold
SalePrice
count  1460.000000   1460.000000  1460.000000  1460.000000
1460.000000
mean      2.758904     43.489041     6.321918  2007.815753
180921.195890
std      40.177307    496.123024     2.703626     1.328095
79442.502883
min       0.000000      0.000000     1.000000  2006.000000
34900.000000
25%       0.000000      0.000000     5.000000  2007.000000
129975.000000
50%       0.000000      0.000000     6.000000  2008.000000
163000.000000
75%       0.000000      0.000000     8.000000  2009.000000
214000.000000
max     738.000000  15500.000000    12.000000  2010.000000
755000.000000

[8 rows x 38 columns]
```

```
df_test.describe()
```

```
                Id   MSSubClass  LotFrontage      LotArea
OverallQual  \
count  1459.000000  1459.000000  1232.000000   1459.000000
1459.000000
mean   2190.000000    57.378341    68.580357   9819.161069
6.078821
```

```
std     421.321334    42.746880    22.376841   4955.517327
1.436812
min    1461.000000    20.000000    21.000000   1470.000000
1.000000
25%    1825.500000    20.000000    58.000000   7391.000000
5.000000
50%    2190.000000    50.000000    67.000000   9399.000000
6.000000
75%    2554.500000    70.000000    80.000000  11517.500000
7.000000
max    2919.000000   190.000000   200.000000  56600.000000
10.000000

       OverallCond    YearBuilt  YearRemodAdd   MasVnrArea
BsmtFinSF1  ...  \
count  1459.000000  1459.000000   1459.000000  1444.000000
1458.000000  ...
mean      5.553804  1971.357779   1983.662783   100.709141
439.203704  ...
std       1.113740    30.390071     21.130467   177.625900
455.268042  ...
min       1.000000  1879.000000   1950.000000     0.000000
0.000000  ...
25%       5.000000  1953.000000   1963.000000     0.000000
0.000000  ...
50%       5.000000  1973.000000   1992.000000     0.000000
350.500000  ...
75%       6.000000  2001.000000   2004.000000   164.000000
753.500000  ...
max       9.000000  2010.000000   2010.000000  1290.000000
4010.000000  ...

        GarageArea   WoodDeckSF  OpenPorchSF  EnclosedPorch
3SsnPorch  \
count  1458.000000  1459.000000  1459.000000    1459.000000
1459.000000
mean    472.768861    93.174777    48.313914      24.243317
1.794380
std     217.048611   127.744882    68.883364      67.227765
20.207842
min       0.000000     0.000000     0.000000       0.000000
0.000000
25%     318.000000     0.000000     0.000000       0.000000
0.000000
50%     480.000000     0.000000    28.000000       0.000000
0.000000
75%     576.000000   168.000000    72.000000       0.000000
0.000000
max    1488.000000  1424.000000   742.000000    1012.000000
360.000000
```

```
       ScreenPorch     PoolArea       MiscVal       MoSold
YrSold
count  1459.000000  1459.000000   1459.000000  1459.000000
1459.000000
mean     17.064428     1.744345     58.167923     6.104181
2007.769705
std      56.609763    30.491646    630.806978     2.722432
1.301740
min       0.000000     0.000000      0.000000     1.000000
2006.000000
25%       0.000000     0.000000      0.000000     4.000000
2007.000000
50%       0.000000     0.000000      0.000000     6.000000
2008.000000
75%       0.000000     0.000000      0.000000     8.000000
2009.000000
max     576.000000   800.000000  17000.000000    12.000000
2010.000000

[8 rows x 37 columns]
```

```
df_train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1460 entries, 0 to 1459
Data columns (total 81 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Id             1460 non-null   int64
 1   MSSubClass     1460 non-null   int64
 2   MSZoning       1460 non-null   object
 3   LotFrontage    1201 non-null   float64
 4   LotArea        1460 non-null   int64
 5   Street         1460 non-null   object
 6   Alley          91 non-null     object
 7   LotShape       1460 non-null   object
 8   LandContour    1460 non-null   object
 9   Utilities      1460 non-null   object
 10  LotConfig      1460 non-null   object
 11  LandSlope      1460 non-null   object
 12  Neighborhood   1460 non-null   object
 13  Condition1     1460 non-null   object
 14  Condition2     1460 non-null   object
 15  BldgType       1460 non-null   object
 16  HouseStyle     1460 non-null   object
 17  OverallQual    1460 non-null   int64
 18  OverallCond    1460 non-null   int64
 19  YearBuilt      1460 non-null   int64
 20  YearRemodAdd   1460 non-null   int64
```

```
21  RoofStyle      1460 non-null   object
22  RoofMatl       1460 non-null   object
23  Exterior1st    1460 non-null   object
24  Exterior2nd    1460 non-null   object
25  MasVnrType     1452 non-null   object
26  MasVnrArea     1452 non-null   float64
27  ExterQual      1460 non-null   object
28  ExterCond      1460 non-null   object
29  Foundation     1460 non-null   object
30  BsmtQual       1423 non-null   object
31  BsmtCond       1423 non-null   object
32  BsmtExposure   1422 non-null   object
33  BsmtFinType1   1423 non-null   object
34  BsmtFinSF1     1460 non-null   int64
35  BsmtFinType2   1422 non-null   object
36  BsmtFinSF2     1460 non-null   int64
37  BsmtUnfSF      1460 non-null   int64
38  TotalBsmtSF    1460 non-null   int64
39  Heating        1460 non-null   object
40  HeatingQC      1460 non-null   object
41  CentralAir     1460 non-null   object
42  Electrical     1459 non-null   object
43  1stFlrSF       1460 non-null   int64
44  2ndFlrSF       1460 non-null   int64
45  LowQualFinSF   1460 non-null   int64
46  GrLivArea      1460 non-null   int64
47  BsmtFullBath   1460 non-null   int64
48  BsmtHalfBath   1460 non-null   int64
49  FullBath       1460 non-null   int64
50  HalfBath       1460 non-null   int64
51  BedroomAbvGr   1460 non-null   int64
52  KitchenAbvGr   1460 non-null   int64
53  KitchenQual    1460 non-null   object
54  TotRmsAbvGrd   1460 non-null   int64
55  Functional     1460 non-null   object
56  Fireplaces     1460 non-null   int64
57  FireplaceQu    770 non-null    object
58  GarageType     1379 non-null   object
59  GarageYrBlt    1379 non-null   float64
60  GarageFinish   1379 non-null   object
61  GarageCars     1460 non-null   int64
62  GarageArea     1460 non-null   int64
63  GarageQual     1379 non-null   object
64  GarageCond     1379 non-null   object
65  PavedDrive     1460 non-null   object
66  WoodDeckSF     1460 non-null   int64
67  OpenPorchSF    1460 non-null   int64
68  EnclosedPorch  1460 non-null   int64
69  3SsnPorch      1460 non-null   int64
70  ScreenPorch    1460 non-null   int64
```

```
 71  PoolArea       1460 non-null   int64
 72  PoolQC         7 non-null      object
 73  Fence          281 non-null    object
 74  MiscFeature    54 non-null     object
 75  MiscVal        1460 non-null   int64
 76  MoSold         1460 non-null   int64
 77  YrSold         1460 non-null   int64
 78  SaleType       1460 non-null   object
 79  SaleCondition  1460 non-null   object
 80  SalePrice      1460 non-null   int64
dtypes: float64(3), int64(35), object(43)
memory usage: 924.0+ KB
```

```
df_test.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1459 entries, 0 to 1458
Data columns (total 80 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Id             1459 non-null   int64
 1   MSSubClass     1459 non-null   int64
 2   MSZoning       1455 non-null   object
 3   LotFrontage    1232 non-null   float64
 4   LotArea        1459 non-null   int64
 5   Street         1459 non-null   object
 6   Alley          107 non-null    object
 7   LotShape       1459 non-null   object
 8   LandContour    1459 non-null   object
 9   Utilities      1457 non-null   object
 10  LotConfig      1459 non-null   object
 11  LandSlope      1459 non-null   object
 12  Neighborhood   1459 non-null   object
 13  Condition1     1459 non-null   object
 14  Condition2     1459 non-null   object
 15  BldgType       1459 non-null   object
 16  HouseStyle     1459 non-null   object
 17  OverallQual    1459 non-null   int64
 18  OverallCond    1459 non-null   int64
 19  YearBuilt      1459 non-null   int64
 20  YearRemodAdd   1459 non-null   int64
 21  RoofStyle      1459 non-null   object
 22  RoofMatl       1459 non-null   object
 23  Exterior1st    1458 non-null   object
 24  Exterior2nd    1458 non-null   object
 25  MasVnrType     1443 non-null   object
 26  MasVnrArea     1444 non-null   float64
 27  ExterQual      1459 non-null   object
 28  ExterCond      1459 non-null   object
 29  Foundation     1459 non-null   object
 30  BsmtQual       1415 non-null   object
```

```
 31  BsmtCond       1414 non-null   object
 32  BsmtExposure   1415 non-null   object
 33  BsmtFinType1   1417 non-null   object
 34  BsmtFinSF1     1458 non-null   float64
 35  BsmtFinType2   1417 non-null   object
 36  BsmtFinSF2     1458 non-null   float64
 37  BsmtUnfSF      1458 non-null   float64
 38  TotalBsmtSF    1458 non-null   float64
 39  Heating        1459 non-null   object
 40  HeatingQC      1459 non-null   object
 41  CentralAir     1459 non-null   object
 42  Electrical     1459 non-null   object
 43  1stFlrSF       1459 non-null   int64
 44  2ndFlrSF       1459 non-null   int64
 45  LowQualFinSF   1459 non-null   int64
 46  GrLivArea      1459 non-null   int64
 47  BsmtFullBath   1457 non-null   float64
 48  BsmtHalfBath   1457 non-null   float64
 49  FullBath       1459 non-null   int64
 50  HalfBath       1459 non-null   int64
 51  BedroomAbvGr   1459 non-null   int64
 52  KitchenAbvGr   1459 non-null   int64
 53  KitchenQual    1458 non-null   object
 54  TotRmsAbvGrd   1459 non-null   int64
 55  Functional     1457 non-null   object
 56  Fireplaces     1459 non-null   int64
 57  FireplaceQu    729 non-null    object
 58  GarageType     1383 non-null   object
 59  GarageYrBlt    1381 non-null   float64
 60  GarageFinish   1381 non-null   object
 61  GarageCars     1458 non-null   float64
 62  GarageArea     1458 non-null   float64
 63  GarageQual     1381 non-null   object
 64  GarageCond     1381 non-null   object
 65  PavedDrive     1459 non-null   object
 66  WoodDeckSF     1459 non-null   int64
 67  OpenPorchSF    1459 non-null   int64
 68  EnclosedPorch  1459 non-null   int64
 69  3SsnPorch      1459 non-null   int64
 70  ScreenPorch    1459 non-null   int64
 71  PoolArea       1459 non-null   int64
 72  PoolQC         3 non-null      object
 73  Fence          290 non-null    object
 74  MiscFeature    51 non-null     object
 75  MiscVal        1459 non-null   int64
 76  MoSold         1459 non-null   int64
 77  YrSold         1459 non-null   int64
 78  SaleType       1458 non-null   object
 79  SaleCondition  1459 non-null   object
```

```
dtypes: float64(11), int64(26), object(43)
memory usage: 912.0+ KB
```

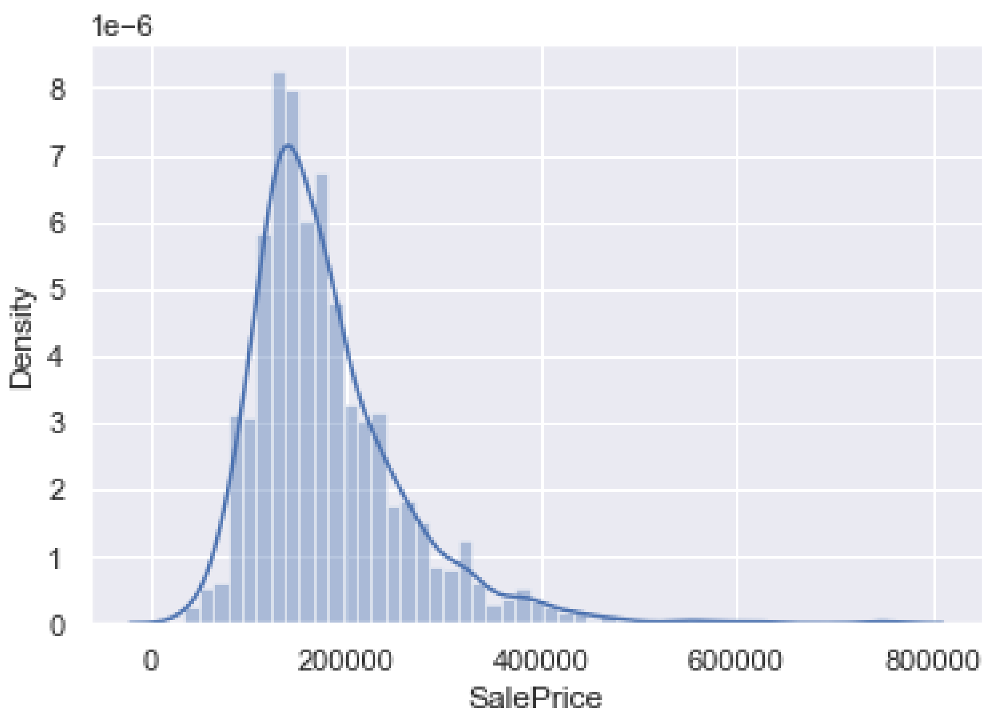
```
df_train['SalePrice'].describe()
```

```
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
```

```
sns.distplot(df_train['SalePrice']);
print("Skewness: %f" % df_train['SalePrice'].skew())
print("Kurtosis: %f" % df_train['SalePrice'].kurt())
```

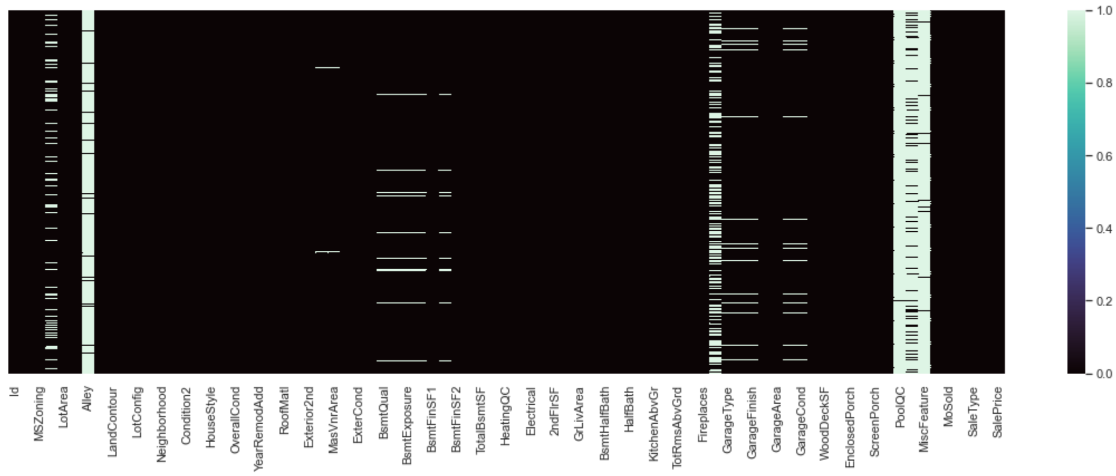
```
C:\Users\16095\anaconda3\lib\site-packages\seaborn\
distributions.py:2557: FutureWarning: `distplot` is a deprecated
function and will be removed in a future version. Please adapt your
code to use either `displot` (a figure-level function with similar
flexibility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```

```
Skewness: 1.882876
Kurtosis: 6.536282
```



```
plt.figure(figsize=(20,6))
sns.heatmap(df_train.isnull(),yticklabels=False,cbar=True,cmap='mako')
```

```
<AxesSubplot:>
```




```
total_null = df_train.isnull().sum().sort_values(ascending=False)
#First sum and order all null values for each variable
percentage =
(df_train.isnull().sum()/df_train.isnull().count()).sort_values(ascend
ing=False) #Get the percentage
missing_data = pd.concat([total_null, percentage], axis=1,
keys=['Total', 'Percentage'])
missing_data.head(20)
```

	Total	Percentage
PoolQC	1453	0.995205
MiscFeature	1406	0.963014
Alley	1369	0.937671
Fence	1179	0.807534
FireplaceQu	690	0.472603
LotFrontage	259	0.177397
GarageYrBlt	81	0.055479
GarageCond	81	0.055479
GarageType	81	0.055479
GarageFinish	81	0.055479
GarageQual	81	0.055479
BsmtFinType2	38	0.026027
BsmtExposure	38	0.026027
BsmtQual	37	0.025342
BsmtCond	37	0.025342
BsmtFinType1	37	0.025342
MasVnrArea	8	0.005479
MasVnrType	8	0.005479
Electrical	1	0.000685
Id	0	0.000000

Categorical

```
categ_vars_ls = ['PoolQC', 'MiscFeature', 'Alley', 'Fence',
                 'FireplaceQu', 'GarageType', 'GarageFinish',
'GarageQual',
                 'GarageCond', 'BsmtQual', 'BsmtCond', 'BsmtExposure',

                 'BsmtFinType1', 'BsmtFinType2', 'MasVnrType']

# Clean train set
for var in categ_vars_ls:
    df_train[var].fillna('None', inplace=True)

# Clean test set
for var in categ_vars_ls:
    df_test[var].fillna('None', inplace=True)
```

Numerical

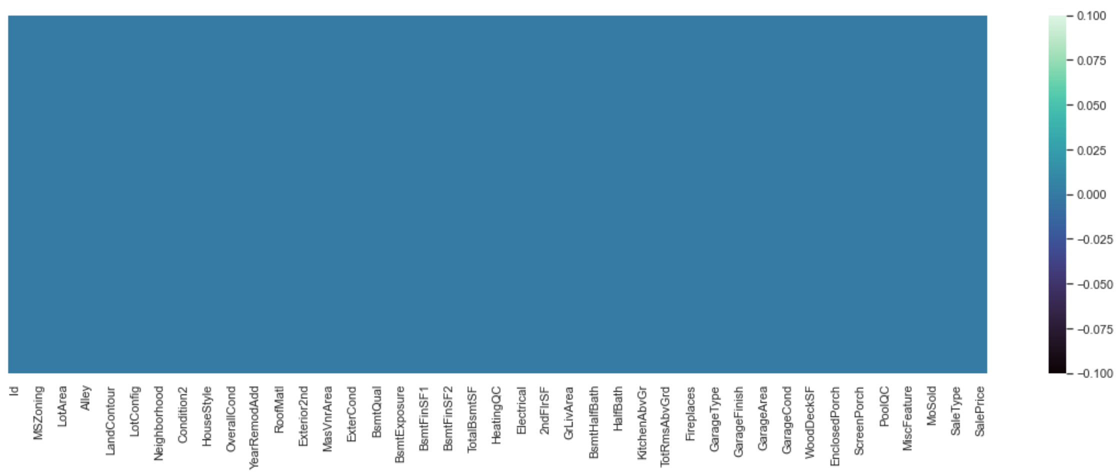
```
num_vars_ls = ['GarageArea', 'GarageCars', 'BsmtFinSF1', 'BsmtFinSF2',

               'BsmtUnfSF', 'TotalBsmtSF', 'BsmtFullBath',
'BsmtHalfBath',
               'MasVnrArea']

# Clean train set
for var in num_vars_ls:
    df_train[var].fillna(0, inplace=True)

# Clean test set
for var in num_vars_ls:
    df_test[var].fillna(0, inplace=True)

vars_ls1 = ['Functional', 'MSZoning', 'Electrical', 'KitchenQual',
'Exterior1st',
            'Exterior2nd', 'SaleType', 'Utilities']

imputer = SimpleImputer(strategy='most_frequent')

# Clean train set
df_train[vars_ls1] =
pd.DataFrame(imputer.fit_transform(df_train[vars_ls1]),
index=df_train.index)

# Clean test set
df_test[vars_ls1] =
pd.DataFrame(imputer.fit_transform(df_test[vars_ls1]),
index=df_test.index)
```

```
train_average_house_neighb = df_train.groupby('Neighborhood')
['LotFrontage']
test_average_house_neighb = df_test.groupby('Neighborhood')
['LotFrontage']

# Clean train set
df_train['LotFrontage'].fillna(train_average_house_neighb.transform(la
mbda x: x.fillna(x.mean())), inplace=True)

# Clean test set
df_test['LotFrontage'].fillna(test_average_house_neighb.transform(lamb
da x: x.fillna(x.mean())), inplace=True)

# Clean train set
df_train['GarageYrBlt'] =
df_train['GarageYrBlt'].fillna(df_train['YearBuilt'])

# Clean test set
df_test['GarageYrBlt'] =
df_test['GarageYrBlt'].fillna(df_test['YearBuilt'])

## NA Check: Verify that we covered all 'NAs' in our data
print(f'Number of NAs in train df: {sum(df_train.isnull().sum())}')
print(f'Number of NAs in test df: {sum(df_test.isnull().sum())}')
```

```
Number of NAs in train df: 0
Number of NAs in test df: 0
```

```
plt.figure(figsize=(20,6))
sns.heatmap(df_train.isnull(),yticklabels=False,cbar=True,cmap='mako')
```

```
<AxesSubplot:>
```



Investigate potential features & outliers

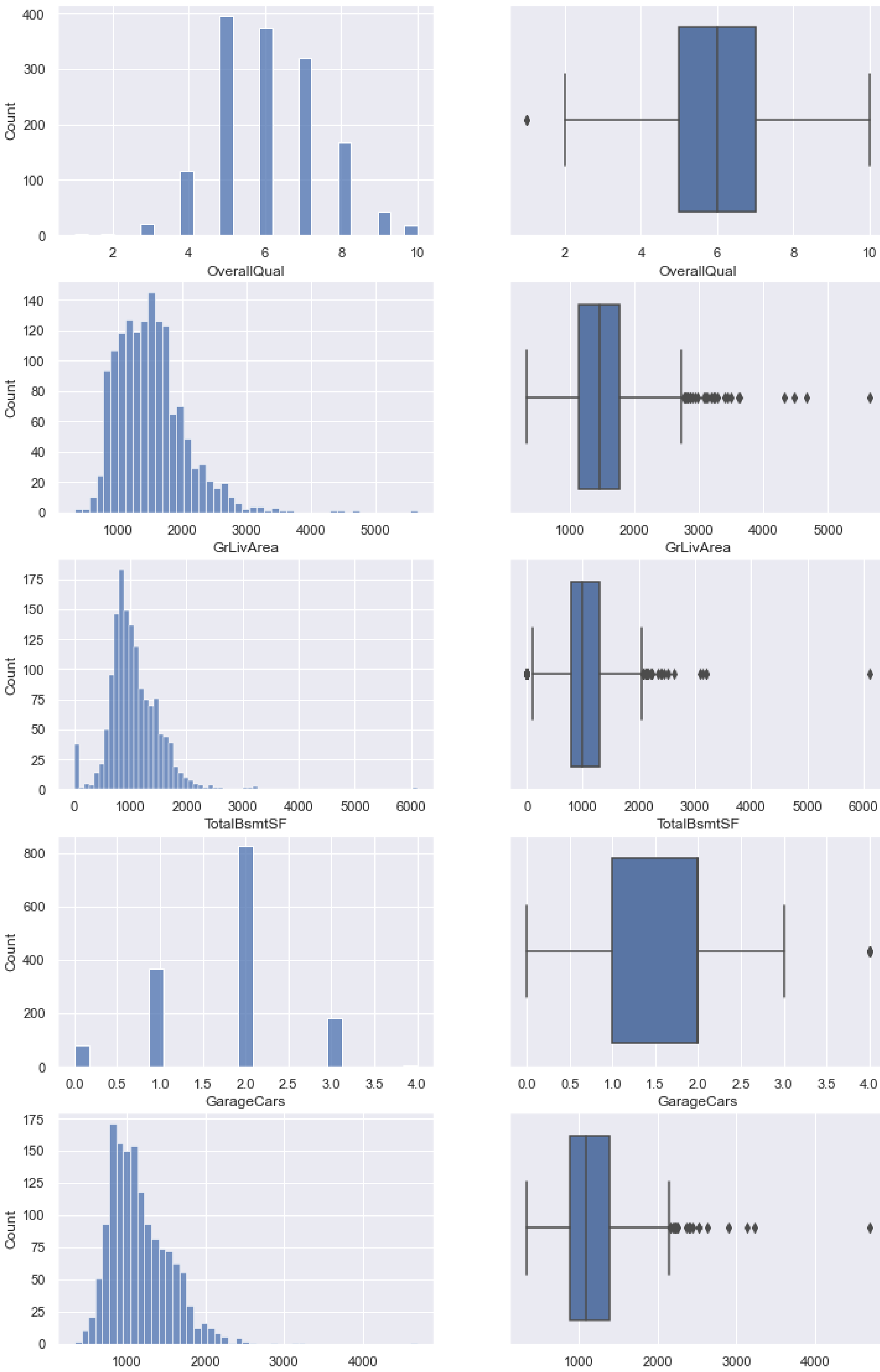
Below, We can see a few of the highest correlating predictors of SalePrice. Based on these features, it is obvious that usable square footage cumulatively amounts to the highest correlation to SalePrice (GrLivArea, TotalBsmtSF, 1stFlrSF, GarageArea). Other discrete and categorical variables (OverallQual, GarageCars, FullBath, TotRmsAdvGrd) influence the dependent variable as well.

```
corr_mat = df_train.corr().SalePrice.sort_values(ascending=False)
corr_mat.head(10)
```

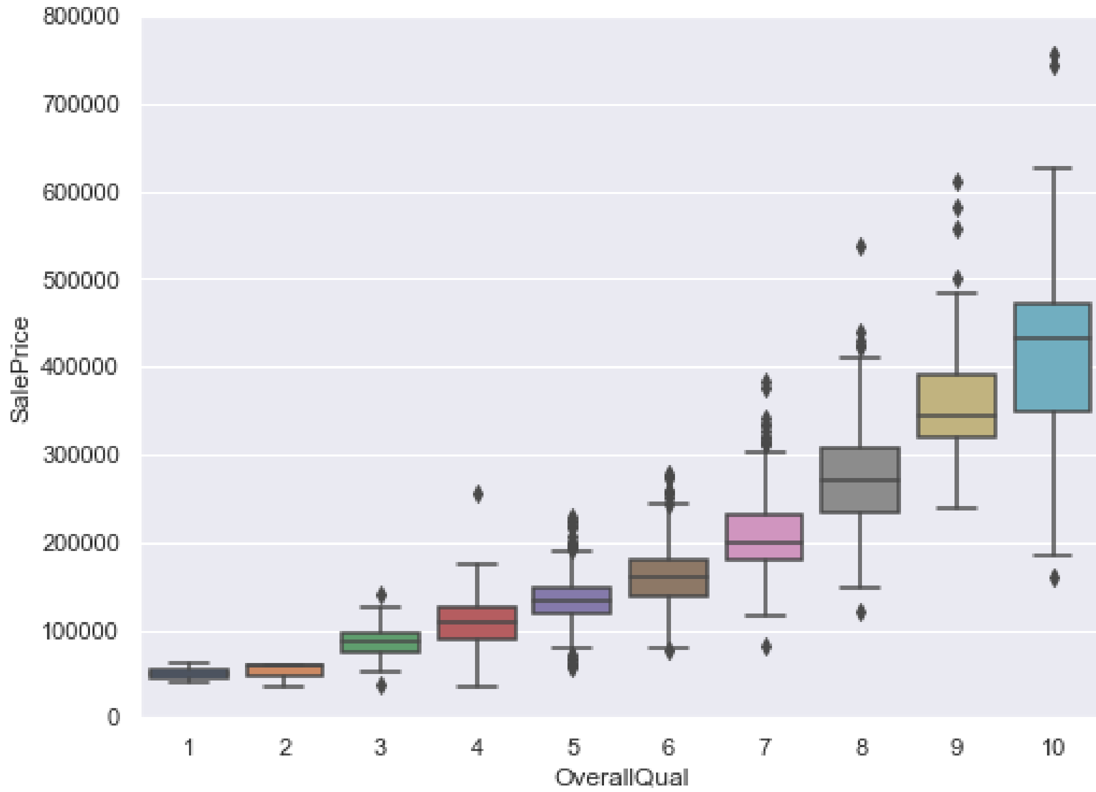
```
SalePrice       1.000000
OverallQual     0.790982
GrLivArea       0.708624
GarageCars      0.640409
GarageArea      0.623431
TotalBsmtSF     0.613581
1stFlrSF        0.605852
FullBath        0.560664
TotRmsAbvGrd    0.533723
YearBuilt       0.522897
Name: SalePrice, dtype: float64
```

Below we can see the distribution of a few of these variables and assess how outliers may impact the data.

```
cor_features = ['OverallQual', 'GrLivArea', 'TotalBsmtSF',
'GarageCars', '1stFlrSF', 'YearBuilt' ]

n = len(cor_features)

fig = plt.figure(figsize=(6*2, 4*n))
# add 2 graph for each column variable
gs = fig.add_gridspec(n, 2)
ax = [[fig.add_subplot(gs[i, j]) for j in range(2)] for i in range(n)]

for i in range(n):
    sns.histplot(x=cor_features[i], data=df_train, ax=ax[i][0])
    sns.boxplot(x=cor_features[i], data=df_train, ax=ax[i][1])

plt.show()
```


Count
OverallQual
OverallQual
Count
GrLivArea
GrLivArea
Count
TotalBsmtSF
TotalBsmtSF
Count
GarageCars
GarageCars
Count
1stFlrSF
1stFlrSF

```python
# OverallQual and SalePrice
data = pd.concat([df_train['SalePrice'], df_train['OverallQual']],
axis=1)
f, ax = plt.subplots(figsize=(8, 6))
fig = sns.boxplot(x='OverallQual', y="SalePrice", data=data)
fig.axis(ymin=0, ymax=800000);
```




```python
# GrLivArea and SalesPrice
sns.set_style('darkgrid')
plt.figure(figsize=(8, 6))
sns.scatterplot(x='GrLivArea', y='SalePrice', data=df_train)
title = plt.title('House Price vs. Above Ground Living Area')
```

The scatter plot above reveals a few outliers where a larger living area is recorded with a low sale price. These outliers can be removed to ensure they do not influence future models.

```
# Clean df_train (GrLiveArea)
outlier = df_train[(df_train.GrLivArea > 4000) & (df_train.SalePrice <
200000)].index
df_train.drop(outlier, axis=0, inplace=True)

# TotalBsmtSF and SalesPrice
sns.set_style('darkgrid')
plt.figure(figsize=(8, 6))
sns.scatterplot(x='TotalBsmtSF', y='SalePrice', data=df_train)
title = plt.title('House Price vs. Basement (sqft)')
```


```
# 1stFlrSF and SalesPrice
sns.set_style('darkgrid')
plt.figure(figsize=(8, 6))
sns.scatterplot(x='1stFlrSF', y='SalePrice', data=df_train)
title = plt.title('House Price vs. First Floor (sqft)')
```

Feature Creation

Feature creation is likely to be a useful approach to finding more potent predictors in this data set. Based on the list of high correlating variables, it is apparent that features representing usable square feet are strong predictors and can be merged to create a stronger predictive feature. Additionally, the current dataframe seems to categorically discriminate based on above or below ground features. Combining some of high correlation variable, both above and below ground, may yield an overall stronger predictor. Finally, YearBuilt showed up a on the bottom of the correlation list with a comparatively low correlation. However, it remains an interesting feature to explore given some obvious and real world implications. Ideally, it would be nice to see in depth how larger renovations might impact the value of older homes. However, the data makes it difficult to define what renovation may have occurred.

Potentially interesting new predictors include:

-Total Square Feet of living Space (Below and Above ground)

-Total Number of Bathrooms (Below and Above Ground)

-Age of House when sold

```
# Total Square Feet Column
df_train['TotalSqft'] = df_train['TotalBsmtSF'] + df_train['1stFlrSF']
```

```
+ df_train['2ndFlrSF']

df_test['TotalSqft'] = df_test['TotalBsmtSF'] + df_test['1stFlrSF'] +
df_test['2ndFlrSF']

# Total Bathrooms Column
df_train['TotalBath'] = df_train['FullBath'] +
df_train['BsmtFullBath'] + 0.5*(df_train['HalfBath'] +
df_train['BsmtHalfBath'])

df_test['TotalBath'] = df_test['FullBath'] + df_test['BsmtFullBath'] +
0.5*(df_test['HalfBath'] + df_test['BsmtHalfBath'])

# Age of House
df_train['HouseAge'] = df_train['YrSold'] - df_train['YearBuilt']

df_test['HouseAge'] = df_test['YrSold'] - df_test['YearBuilt']

# Check for new columns
df_train.head()
```

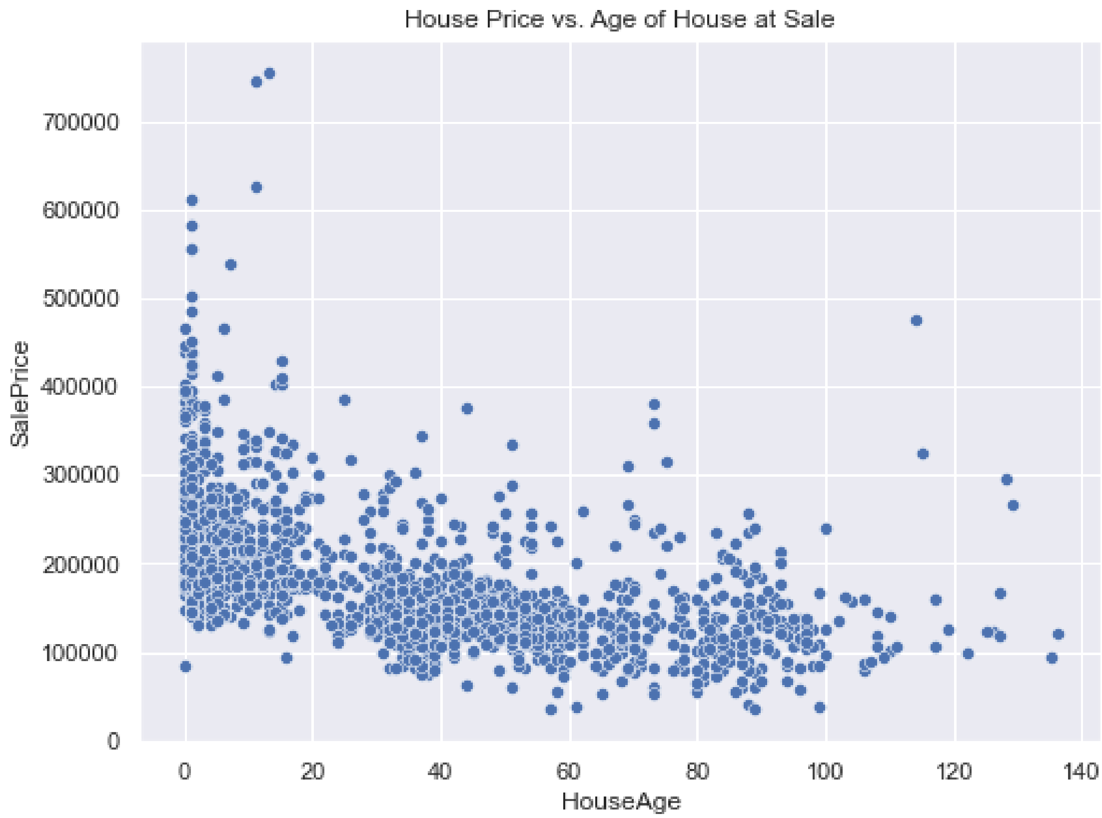
```
   Id  MSSubClass MSZoning  LotFrontage  LotArea Street Alley LotShape
\
0   1          60       RL         65.0     8450   Pave  None      Reg

1   2          20       RL         80.0     9600   Pave  None      Reg

2   3          60       RL         68.0    11250   Pave  None      IR1

3   4          70       RL         60.0     9550   Pave  None      IR1

4   5          60       RL         84.0    14260   Pave  None      IR1


  LandContour Utilities  ... MiscFeature MiscVal MoSold YrSold
SaleType  \
0         Lvl    AllPub  ...        None       0      2   2008
WD
1         Lvl    AllPub  ...        None       0      5   2007
WD
2         Lvl    AllPub  ...        None       0      9   2008
WD
3         Lvl    AllPub  ...        None       0      2   2006
WD
4         Lvl    AllPub  ...        None       0     12   2008
WD

  SaleCondition SalePrice  TotalSqft  TotalBath  HouseAge
0        Normal    208500       2566        3.5         5
1        Normal    181500       2524        2.5        31
```

```
2         Normal    223500       2706        3.5          7
3        Abnorml    140000       2473        2.0         91
4         Normal    250000       3343        3.5          8

[5 rows x 84 columns]
```

```
# TotalSqft and SalesPrice
sns.set_style('darkgrid')
plt.figure(figsize=(8, 6))
sns.scatterplot(x='TotalSqft', y='SalePrice', data=df_train)
title = plt.title('House Price vs. Total Living Space')
```


```
# TotalBath and SalesPrice
sns.set_style('darkgrid')
plt.figure(figsize=(8, 6))
sns.scatterplot(x='TotalBath', y='SalePrice', data=df_train)
title = plt.title('House Price vs. Total Bathrooms')
```


```
# HouseAge and SalesPrice
sns.set_style('darkgrid')
plt.figure(figsize=(8, 6))
sns.scatterplot(x='HouseAge', y='SalePrice', data=df_train)
title = plt.title('House Price vs. Age of House at Sale')
```




```
corr_mat2 = df_train.corr().SalePrice.sort_values(ascending=False)
corr_mat2.head(10)
```

```
SalePrice      1.000000
TotalSqft      0.832877
OverallQual    0.795774
GrLivArea      0.734968
TotalBsmtSF    0.651153
GarageCars     0.641047
TotalBath      0.635896
1stFlrSF       0.631530
GarageArea     0.629217
FullBath       0.562165
Name: SalePrice, dtype: float64
```

Prep

```
corr_mat = df_train.corr()
corr_mat['SalePrice'][(corr_mat["SalePrice"] > 0.50)]
```

```
OverallQual     0.795774
YearBuilt       0.523608
YearRemodAdd    0.507717
TotalBsmtSF     0.651153
1stFlrSF        0.631530
```

```
GrLivArea       0.734968
FullBath        0.562165
TotRmsAbvGrd    0.537769
GarageYrBlt     0.508719
GarageCars      0.641047
GarageArea      0.629217
SalePrice       1.000000
TotalSqft       0.832877
TotalBath       0.635896
Name: SalePrice, dtype: float64
```

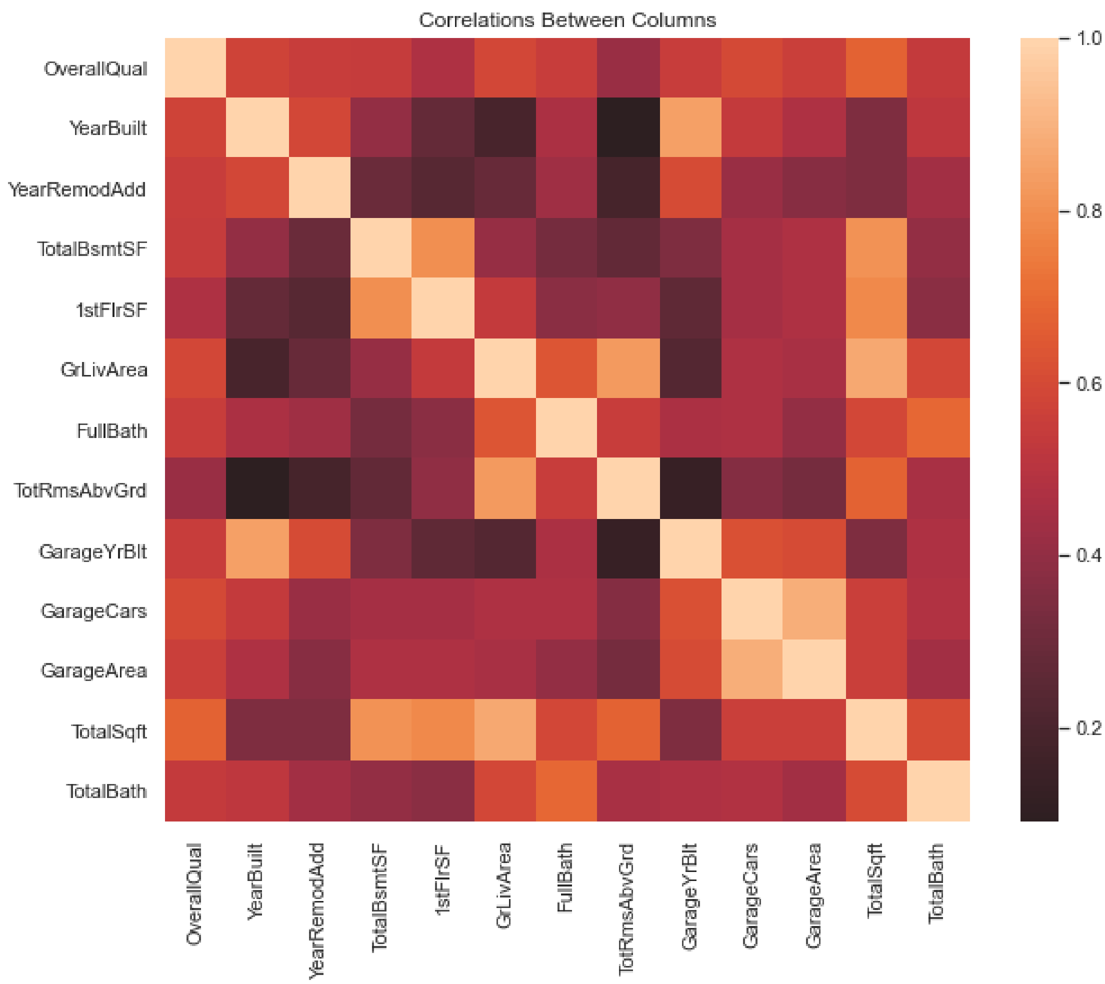
```
important_num_cols = list(corr_mat['SalePrice'][(corr_mat["SalePrice"]
> 0.5)].index)

important_num_cols.remove('SalePrice')
len(important_num_cols)
print(important_num_cols)
```

```
['OverallQual', 'YearBuilt', 'YearRemodAdd', 'TotalBsmtSF',
'1stFlrSF', 'GrLivArea', 'FullBath', 'TotRmsAbvGrd', 'GarageYrBlt',
'GarageCars', 'GarageArea', 'TotalSqft', 'TotalBath']
```

```
x_num_only = df_train[important_num_cols]
x_num_only.shape
```

```
(1458, 13)
```

```
plt.figure(figsize=(10,8))
sns.heatmap(x_num_only.corr(), center = 0)
plt.title("Correlations Between Columns")
plt.show()
```




```
corr_x = x_num_only.corr()

for i in range(0, len(corr_x) - 1):
    for j in range(i + 1, len(corr_x)):
        if(corr_x.iloc[i, j] < -0.6 or corr_x.iloc[i, j] > 0.6):
            print(f"corr: {corr_x.iloc[i, j]}, row: {i}, column: {j};
{corr_x.index[i]}, {corr_x.index[j]}")
```

```
corr: 0.6007408233586199, row: 0, column: 9; OverallQual, GarageCars
corr: 0.6773246628152239, row: 0, column: 11; OverallQual, TotalSqft
corr: 0.8448398583188307, row: 1, column: 8; YearBuilt, GarageYrBlt
corr: 0.6036359426335939, row: 2, column: 8; YearRemodAdd, GarageYrBlt
corr: 0.8038296279256135, row: 3, column: 4; TotalBsmtSF, 1stFlrSF
corr: 0.8063997413782633, row: 3, column: 11; TotalBsmtSF, TotalSqft
corr: 0.7819168108101016, row: 4, column: 11; 1stFlrSF, TotalSqft
corr: 0.6383784637415104, row: 5, column: 6; GrLivArea, FullBath
corr: 0.8294981976715332, row: 5, column: 7; GrLivArea, TotRmsAbvGrd
corr: 0.8663861141668201, row: 5, column: 11; GrLivArea, TotalSqft
corr: 0.6932148078878585, row: 6, column: 12; FullBath, TotalBath
corr: 0.6785607174896137, row: 7, column: 11; TotRmsAbvGrd, TotalSqft
corr: 0.6195177230785245, row: 8, column: 9; GarageYrBlt, GarageCars
```

```
corr: 0.6030387164134529, row: 8, column: 10; GarageYrBlt, GarageArea
corr: 0.8873044983919188, row: 9, column: 10; GarageCars, GarageArea
corr: 0.608161005563307, row: 11, column: 12; TotalSqft, TotalBath
```

```
# from the information above, we want to drop '1stFlrSF', 'FullBath',
'TotRmsAbvGrd', and 'GarageArea'
num_cols_ls = [i for i in x_num_only.columns if i not in ['1stFlrSF',
'FullBath', 'TotRmsAbvGrd', 'GarageArea']]
num_cols_ls
```

```
['OverallQual',
 'YearBuilt',
 'YearRemodAdd',
 'TotalBsmtSF',
 'GrLivArea',
 'GarageYrBlt',
 'GarageCars',
 'TotalSqft',
 'TotalBath']
```

```
# Select the important categorical features to use
cat_cols = ["MSZoning", "Utilities", "BldgType", "Heating",
            "KitchenQual", "SaleCondition", "LandSlope"]

columns_full_ls = num_cols_ls + cat_cols
columns_full_ls
```

```
['OverallQual',
 'YearBuilt',
 'YearRemodAdd',
 'TotalBsmtSF',
 'GrLivArea',
 'GarageYrBlt',
 'GarageCars',
 'TotalSqft',
 'TotalBath',
 'MSZoning',
 'Utilities',
 'BldgType',
 'Heating',
 'KitchenQual',
 'SaleCondition',
 'LandSlope']
```

```
df_train_final = df_train
df_test_final = df_test

df_train_final.head()
```

```
   Id  MSSubClass MSZoning  LotFrontage  LotArea Street Alley LotShape
\
```

```
0   1           60       RL          65.0     8450   Pave  None      Reg

1   2           20       RL          80.0     9600   Pave  None      Reg

2   3           60       RL          68.0    11250   Pave  None      IR1

3   4           70       RL          60.0     9550   Pave  None      IR1

4   5           60       RL          84.0    14260   Pave  None      IR1


  LandContour Utilities  ... MiscFeature MiscVal MoSold YrSold
SaleType  \
0         Lvl    AllPub  ...        None       0      2   2008
WD
1         Lvl    AllPub  ...        None       0      5   2007
WD
2         Lvl    AllPub  ...        None       0      9   2008
WD
3         Lvl    AllPub  ...        None       0      2   2006
WD
4         Lvl    AllPub  ...        None       0     12   2008
WD

  SaleCondition SalePrice  TotalSqft  TotalBath  HouseAge
0        Normal    208500       2566        3.5         5
1        Normal    181500       2524        2.5        31
2        Normal    223500       2706        3.5         7
3       Abnorml    140000       2473        2.0        91
4        Normal    250000       3343        3.5         8

[5 rows x 84 columns]
```

```
for col in columns_full_ls:
    train = sorted(df_train_final[col].unique().tolist())
    test = sorted(df_test_final[col].unique().tolist())
    total = set(train + test)
    df_train_final[col] = pd.Categorical(df_train_final[col],
categories=total)
    df_test_final[col] = pd.Categorical(df_test_final[col],
categories=total)

df_train_final = pd.get_dummies(df_train_final, columns=cat_cols)
df_test_final = pd.get_dummies(df_test_final, columns=cat_cols)


df_train_final.head()
```

```
   Id  MSSubClass  LotFrontage  LotArea Street Alley LotShape 
LandContour  \
0   1          60         65.0     8450   Pave  None      Reg         
Lvl   
1   2          20         80.0     9600   Pave  None      Reg         
Lvl   
2   3          60         68.0    11250   Pave  None      IR1         
Lvl   
3   4          70         60.0     9550   Pave  None      IR1         
Lvl   
4   5          60         84.0    14260   Pave  None      IR1         
Lvl   

  LotConfig Neighborhood  ... KitchenQual_TA SaleCondition_Family  \
0    Inside      CollgCr  ...              0                    0   
1       FR2      Veenker  ...              1                    0   
2    Inside      CollgCr  ...              0                    0   
3    Corner      Crawfor  ...              0                    0   
4       FR2      NoRidge  ...              0                    0   

  SaleCondition_Abnorml SaleCondition_Normal  SaleCondition_AdjLand  \
0                     0                    1                      0   
1                     0                    1                      0   
2                     0                    1                      0   
3                     1                    0                      0   
4                     0                    1                      0   

  SaleCondition_Partial SaleCondition_Alloca LandSlope_Sev 
LandSlope_Mod  \
0                     0                    0             0             
0   
1                     0                    0             0             
0   
2                     0                    0             0             
0   
3                     0                    0             0             
0   
4                     0                    0             0             
0   

  LandSlope_Gtl  
0             1  
1             1  
2             1  
3             1  
4             1  

[5 rows x 108 columns]
```

Simple linear Regression

```
# Feature(s) to look at
f1 = ['TotalSqft']

# Run a Linear Regression using the feature(s)
x1 = df_train_final[f1]
y = df_train_final['SalePrice']

# Split the data
x_train, x_test, y_train, y_test = train_test_split(x1, y)
x_train.shape, x_test.shape, y_train.shape, y_test.shape
```

```
((1093, 1), (365, 1), (1093,), (365,))
```

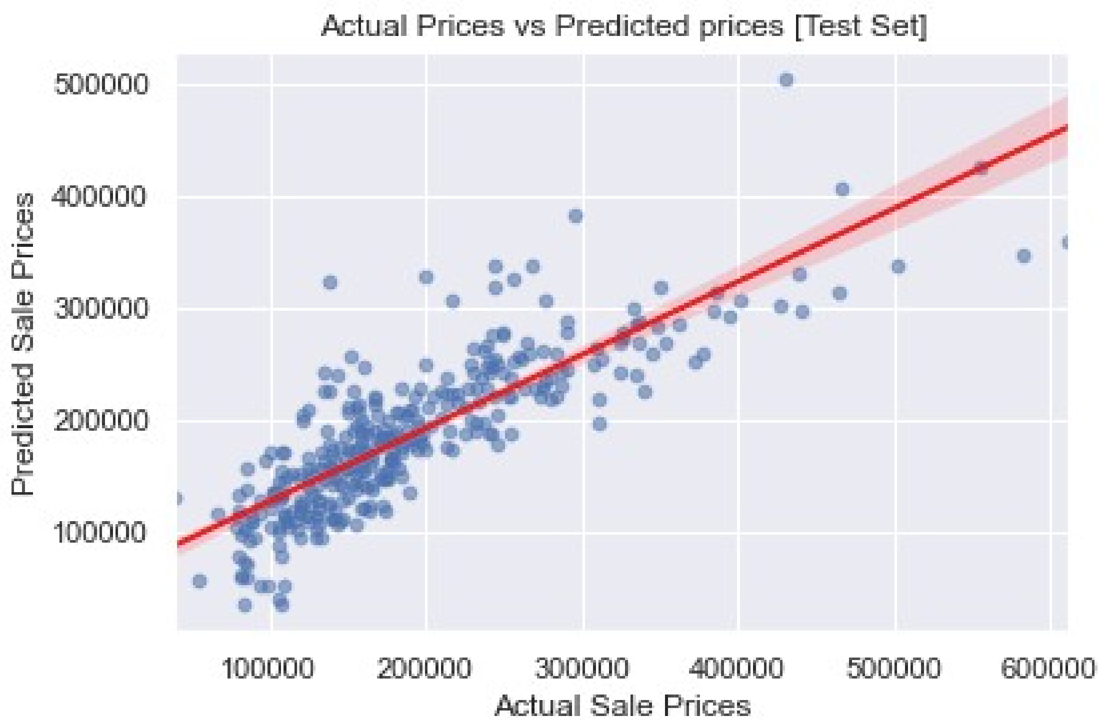
```
# Set up model
linreg1 = LinearRegression()
kf = KFold(n_splits=7, shuffle=True)

# Standardize the data
ss = StandardScaler()
ss_train = ss.fit_transform(x_train)
ss_test = ss.transform(x_test)

# cross validate
scores = cross_val_score(linreg1, ss_train, y_train, cv=kf)
print(scores)
print(f'Mean Score: {scores.mean()}; SD: {scores.std()}')


linreg1.fit(ss_train, y_train)
print(f'TRAIN Score: {linreg1.score(ss_train, y_train)}')
print(f'TEST Score: {linreg1.score(ss_test, y_test)}')

pred = linreg1.predict(ss_test)
b, m = np.polynomial.polynomial.polyfit(y_test, pred, 1)
```

```
[0.62748016 0.77416605 0.62688057 0.64889819 0.69084112 0.73128427
 0.66161388]
Mean Score: 0.6801663223202056; SD: 0.05140877822295907
TRAIN Score: 0.6925391178489992
TEST Score: 0.6925774318501976
```

The mean score is .68 which is not as good as it could be. The training score is not indicate overfitting.

```
# Visualize the model results
sns.scatterplot(x=y_test, y=pred, alpha=0.4)
sns.regplot(x=y_test, y=pred, truncate=True, scatter_kws={'s': 20,
'alpha':0.3},
            line_kws={'color':'red', 'linewidth': 2})
```

```
sns.lineplot(x=np.unique(y_test), y=np.unique(np.poly1d(b + m *
np.unique(y_test))), linewidth=0.5, color='r')

plt.xlabel("Actual Sale Prices")
plt.ylabel("Predicted Sale Prices")
plt.title("Actual Prices vs Predicted prices [Test Set]")

plt.show()
```



Multiple Linear Regression

```
# Feature(s) to look at
f2 = num_cols_ls

# Run a Linear Regression using the feature(s)
x2 = df_train_final[f2]
y = df_train_final['SalePrice']

# Split the data
x2_train, x2_test, y2_train, y2_test = train_test_split(x2, y)
x2_train.shape, x2_test.shape, y2_train.shape, y2_test.shape
```

```
((1093, 9), (365, 9), (1093,), (365,))
```

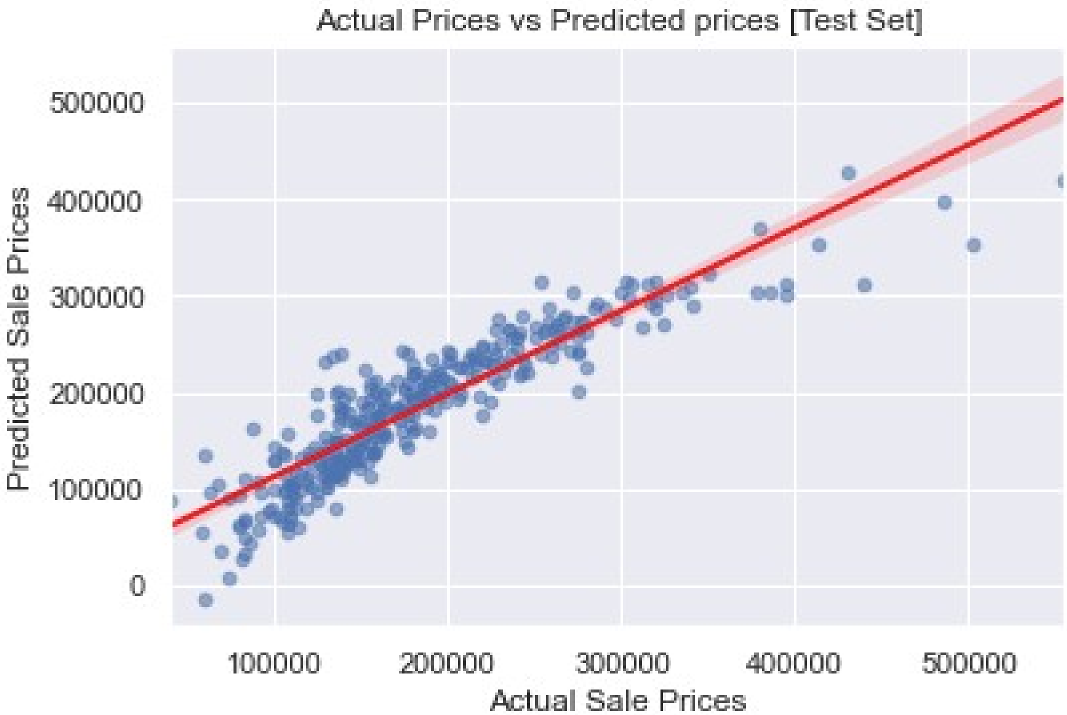
```
# Multi Linear Regression
linreg2 = LinearRegression()
kf = KFold(n_splits=7, shuffle=True)
```

```python
# Standardize the numeric columns
ss = StandardScaler()

ss_train2 = ss.fit_transform(x2_train)
ss_test2 = ss.transform(x2_test)

scores = cross_val_score(linreg2, ss_train2, y2_train, cv=kf)
print(scores)
print(f'Mean Score: {scores.mean()}; SD: {scores.std()}')


linreg2.fit(ss_train2, y2_train)
print(f'TRAIN Score: {linreg2.score(ss_train2, y2_train)}')
print(f'TEST Score: {linreg2.score(ss_test2, y2_test)}')

pred2 = linreg2.predict(ss_test2)
b, m = np.polynomial.polynomial.polyfit(y2_test, pred2, 1)
```

```
[0.84800476 0.7755963  0.81937807 0.81932924 0.82138407 0.77017009
 0.80274092]
Mean Score: 0.8080862058622865; SD: 0.025483983264919123
TRAIN Score: 0.8159443440981854
TEST Score: 0.8186095109797409
```

Based on the scores above, the Multiple Regression model seems to have performed better than the Simple Regression model. It appears to be fitting to the data better than the previous model, but does not indicate overfitting.

```python
sns.scatterplot(x=y2_test, y=pred2, alpha=0.4)
sns.regplot(x=y2_test, y=pred2, truncate=True, scatter_kws={'s': 20,
'alpha':0.3},
            line_kws={'color':'red', 'linewidth': 2})
sns.lineplot(x=np.unique(y2_test), y=np.unique(np.poly1d(b + m *
np.unique(y2_test))), linewidth=0.5, color='r')

plt.xlabel("Actual Sale Prices")
plt.ylabel("Predicted Sale Prices")
plt.title("Actual Prices vs Predicted prices [Test Set]")

plt.show()
```



Conclusion

I am not entirely certain as to why the multiple regression performed better than simple regression in this case. Perhaps the entirety of the numerical list allowed the model to identify more correlative patterns which in turn allowed for a better fit. I expected that the TotalSqft variable would have provided a better fit considering it encompassed several highly correlative variables in one feature. Nevertheless, both models provide a good enough fit to reflect the inputted data.