In [1]:

```
import pandas as pd
```

In [32]:

```
d='s3://datasciencebuckett/train/train-1 (1).csv'
df=pd.read_csv(d)
df
```

Out[32]:

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

1460 rows × 81 columns

Data Cleaning

In [9]:

```
df.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
```

In [10]:

```
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
sns.set_style('whitegrid')
```

In [11]:

```
df.shape
```

Out[11]:

```
(1460, 81)
```

In [12]:

```
print(df.SalePrice.value_counts())
df['SalePrice'].value_counts(normalize=True)
```

```
140000    20
135000    17
145000    14
155000    14
190000    13
          ..
84900      1
424870     1
415298     1
62383      1
34900      1
Name: SalePrice, Length: 663, dtype: int64
```

Out[12]:

```
140000    0.013699
135000    0.011644
145000    0.009589
155000    0.009589
190000    0.008904
            ...
84900     0.000685
424870    0.000685
415298    0.000685
62383     0.000685
34900     0.000685
Name: SalePrice, Length: 663, dtype: float64
```

In [13]:

```
df.isnull().sum()
```

Out[13]:

```
Id                 0
MSSubClass         0
MSZoning           0
LotFrontage      259
LotArea            0
                ...
MoSold             0
YrSold             0
SaleType           0
SaleCondition      0
SalePrice          0
Length: 81, dtype: int64
```

In [14]:

```
df['Alley'].isnull().sum()
```

Out[14]:

```
1369
```

In [15]:

```
df.describe()
```

Out[15]:

	Id	MSSubClass	LotFrontage	LotArea	OverallQual	OverallCond	Ye
count	1460.000000	1460.000000	1201.000000	1460.000000	1460.000000	1460.000000	1460.(
mean	730.500000	56.897260	70.049958	10516.828082	6.099315	5.575342	1971.:
std	421.610009	42.300571	24.284752	9981.264932	1.382997	1.112799	30.:
min	1.000000	20.000000	21.000000	1300.000000	1.000000	1.000000	1872.(
25%	365.750000	20.000000	59.000000	7553.500000	5.000000	5.000000	1954.(
50%	730.500000	50.000000	69.000000	9478.500000	6.000000	5.000000	1973.(
75%	1095.250000	70.000000	80.000000	11601.500000	7.000000	6.000000	2000.(
max	1460.000000	190.000000	313.000000	215245.000000	10.000000	9.000000	2010.(

8 rows × 38 columns

In [16]:

```

```

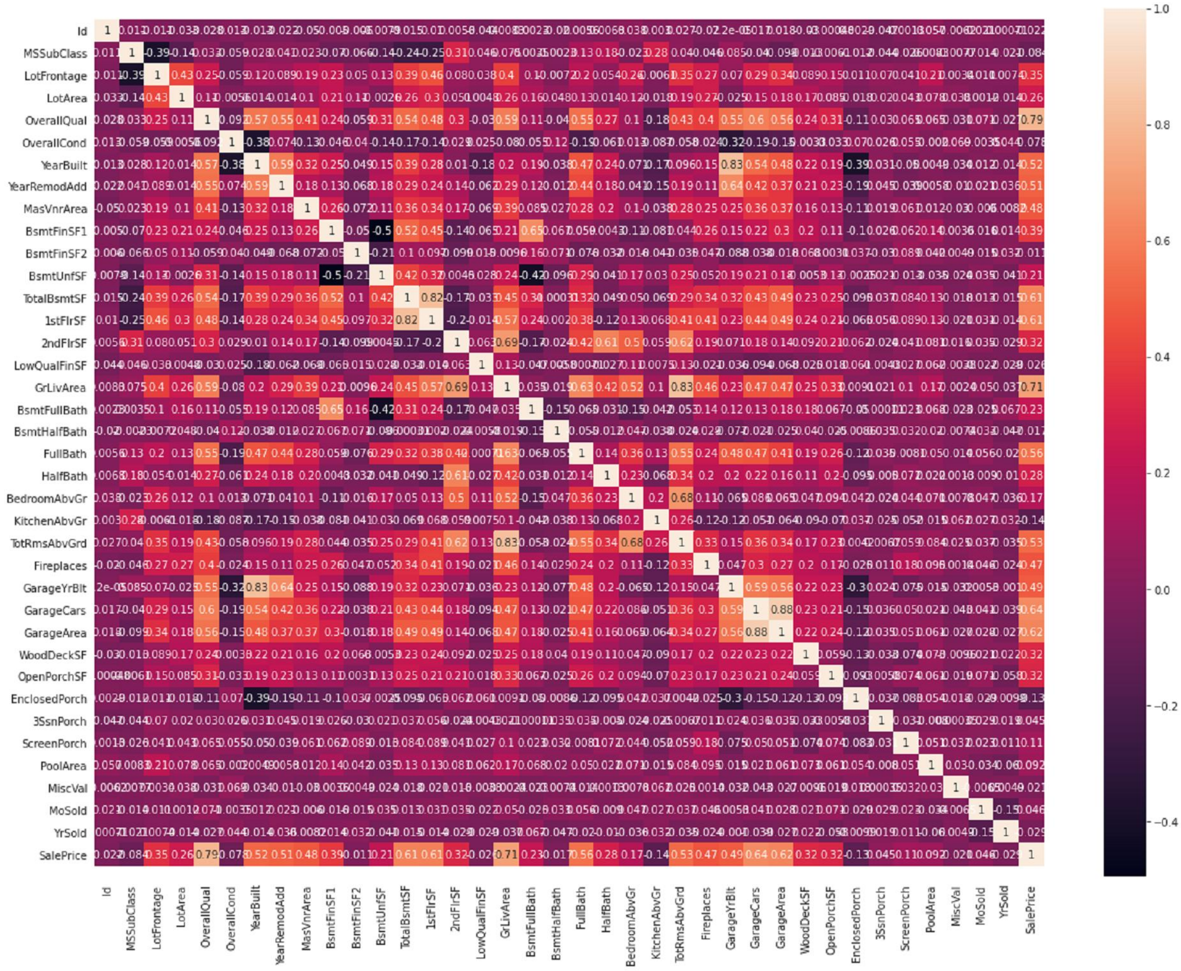
Out[16]:

```
<bound method IndexOpsMixin.value_counts of 0       Pave
1       Pave
2       Pave
3       Pave
4       Pave
        ...
1455    Pave
1456    Pave
1457    Pave
1458    Pave
1459    Pave
Name: Street, Length: 1460, dtype: object>
```

In [20]:

```
#Heatmap to show the correlation between various variables of the dataset

plt.figure(figsize=(20, 15))
cor = df.corr()
ax = sns.heatmap(cor,annot=True)
bottom, top = ax.get_ylim()
ax.set_ylim(bottom + 0.5, top - 0.5)
plt.show()
```

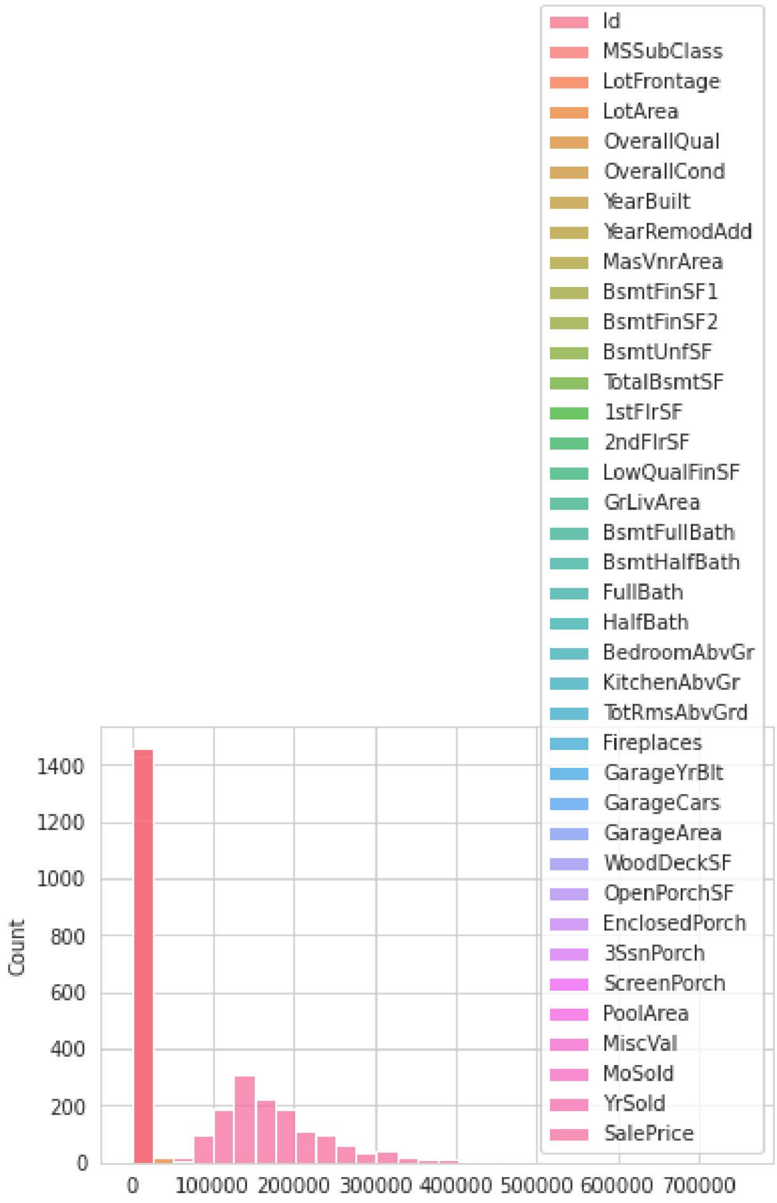


In [21]:

```
sns.histplot(data=df,bins=30)
```

Out[21]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f1461b5ba90>
```



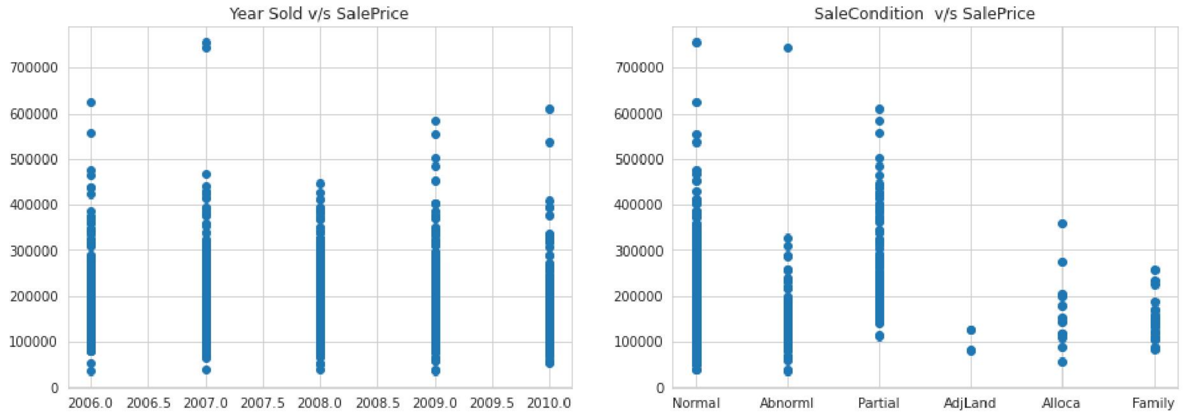
In [23]:

```
fig, (ax1, ax2) = plt.subplots(1,2,figsize = (15,5))

#scatter plot 1
ax1.scatter(x=df['YrSold'],y= df['SalePrice'])
ax1.set_title('Year Sold v/s SalePrice')

#scatter plot 2
ax2.scatter(x=df['SaleCondition'],y=df['SalePrice'])
ax2.set_title('SaleCondition  v/s SalePrice')

plt.draw()
```



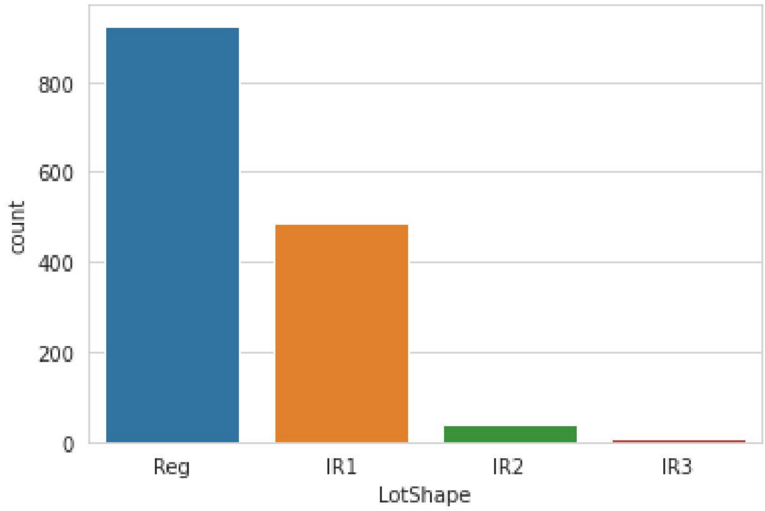
In [27]:

```
sns.countplot(df['LotShape'])
```

```
/home/ec2-user/anaconda3/envs/tensorflow_p36/lib/python3.6/site-packages/sea
born/_decorators.py:43: FutureWarning: Pass the following variable as a keyw
ord arg: x. From version 0.12, the only valid positional argument will be `d
ata`, and passing other arguments without an explicit keyword will result in
an error or misinterpretation.
  FutureWarning
```

Out[27]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f146170b5f8>
```



In [28]:

```
df.isna().any()[lambda x: x]
```

Out[28]:

```
LotFrontage     True
Alley           True
MasVnrType      True
MasVnrArea      True
BsmtQual        True
BsmtCond        True
BsmtExposure    True
BsmtFinType1    True
BsmtFinType2    True
Electrical      True
FireplaceQu     True
GarageType      True
GarageYrBlt     True
GarageFinish    True
GarageQual      True
GarageCond      True
PoolQC          True
Fence           True
MiscFeature     True
dtype: bool
```

In [29]:

```
df[df.columns[df.isnull().any()]]
```

Out[29]:

	LotFrontage	Alley	MasVnrType	MasVnrArea	BsmtQual	BsmtCond	BsmtExposure	Bsm
0	65.0	NaN	BrkFace	196.0	Gd	TA	No	
1	80.0	NaN	None	0.0	Gd	TA	Gd	
2	68.0	NaN	BrkFace	162.0	Gd	TA	Mn	
3	60.0	NaN	None	0.0	TA	Gd	No	
4	84.0	NaN	BrkFace	350.0	Gd	TA	Av	
...	...	...	...	...	...	...	...	
1455	62.0	NaN	None	0.0	Gd	TA	No	
1456	85.0	NaN	Stone	119.0	Gd	TA	No	
1457	66.0	NaN	None	0.0	TA	Gd	No	
1458	68.0	NaN	None	0.0	TA	TA	Mn	
1459	75.0	NaN	None	0.0	TA	TA	No	

1460 rows × 19 columns

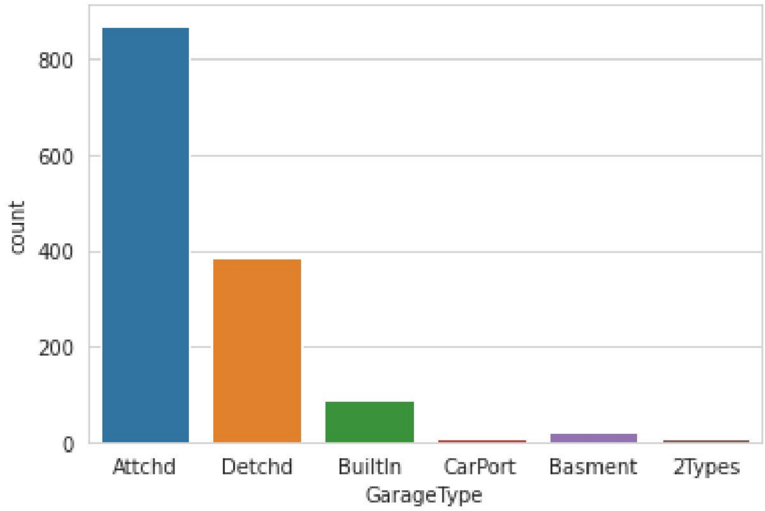
In [30]:

```
sns.countplot(df['GarageType'])
```

```
/home/ec2-user/anaconda3/envs/tensorflow_p36/lib/python3.6/site-packages/sea
born/_decorators.py:43: FutureWarning: Pass the following variable as a keyw
ord arg: x. From version 0.12, the only valid positional argument will be `d
ata`, and passing other arguments without an explicit keyword will result in
an error or misinterpretation.
  FutureWarning
```

Out[30]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f1460eeb5f8>
```



In [34]:

```
df=df.drop(['Alley','PoolQC','Fence','MiscFeature'],axis=1)
```

In [35]:

```
df.head()
```

Out[35]:

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

5 rows × 77 columns

In [36]:

```
df1=pd.get_dummies(df)
df1
```

Out[36]:

	Id	MSSubClass	LotFrontage	LotArea	OverallQual	OverallCond	YearBuilt	YearRemo
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	
...	...	...	...	...	...	...	...	
1455	1456	60	62.0	7917	6	5	1999	
1456	1457	20	85.0	13175	6	6	1978	
1457	1458	70	66.0	9042	7	9	1941	
1458	1459	20	68.0	9717	5	6	1950	
1459	1460	20	75.0	9937	5	6	1965	

1460 rows × 277 columns

In [43]:

```
X=df1.dropna(axis=1)
y=df['SalePrice']
```

Train and Test

In [44]:

```
from sklearn.model_selection import train_test_split

X_train , X_test , y_train , y_test = train_test_split(X,y,test_size = 0.30 , random_state

print(X_train.shape)
print(X_test.shape)
print(y_train.shape)
print(y_test.shape)
```

```
(1022, 274)
(438, 274)
(1022,)
(438,)
```

In [45]:

```
from sklearn.linear_model import LinearRegression
lm = LinearRegression()
lm.fit(X_train,y_train)
```

Out[45]:

```
LinearRegression()
```

In [47]:

```
#The coefficients
print('Coefficients: \n', lm.coef_)
```

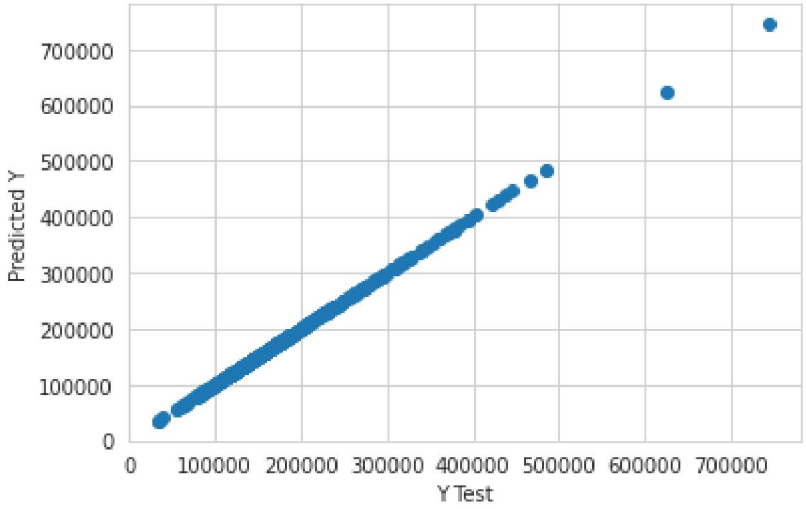
```
Coefficients: 
 [-4.98582163e-15 -8.06243960e-13 -1.11022302e-15  4.79456579e-12
 -4.97977787e-12 -1.56953443e-13 -5.60857133e-13 -1.55785758e-14
  8.58155777e-14 -5.32250025e-14  1.66571604e-14 -4.67118007e-14
 -5.50872282e-14  5.37727957e-14 -4.74373145e-14 -4.71437277e-11
 -3.48604621e-11 -7.07572876e-12  1.25995242e-11  2.92595807e-11
  3.82842738e-11 -9.87925841e-12 -1.50223882e-12 -3.22626736e-11
  1.02175439e-13  2.40262759e-14 -1.44132295e-13  1.04588473e-14
 -1.37817876e-14 -1.37354719e-13 -2.54463830e-13  1.09618750e-14
  2.37663498e-12  4.30381878e-12  1.00000000e+00  2.17211393e-11
 -4.74760340e-11  1.95738782e-11 -4.89274619e-12  1.10737628e-11
 -4.64233051e-12  4.64233050e-12 -9.48371851e-12 -1.49651025e-11
  5.46745327e-11 -3.02257117e-11  2.35150986e-12 -1.21105326e-11
  3.14018950e-11 -2.16428723e-11 -1.41123439e-11  1.41123439e-11
 -1.31419164e-11  4.03032033e-12 -1.27961798e-11  3.63126340e-11
 -1.44048583e-11 -1.76512694e-11 -1.79812709e-11  3.56325404e-11
  4.60689555e-12  9.21089457e-12 -3.43873650e-11  3.44519246e-11
  2.32313555e-12  5.78115685e-13 -3.62677162e-12  1.19927749e-11
  2.94189558e-11 -2.98331551e-11 -3.38111605e-11  4.70457674e-12
  1.80533330e-13  9.29138413e-12  1.07897771e-11  2.50219333e-11
 -6.11446033e-12 -1.48369286e-11 -2.14351782e-11  7.86026136e-12
  2.62074782e-11  3.98497527e-11 -2.86502479e-11  7.24789292e-12
 -5.10410193e-11  1.02611451e-11 -7.84318855e-13 -1.51197561e-11
 -6.59240352e-12 -8.64073920e-12  3.74253239e-13 -1.23579029e-12
 -7.76421442e-12  2.95018240e-11  2.88535184e-11 -9.98315771e-12
  3.44339053e-11 -4.54556950e-11  9.17276500e-12  8.23549007e-12
  4.23516474e-22 -2.52568261e-11 -2.33631248e-11 -9.75523523e-12
  1.02992213e-11  1.55814877e-11  7.23765112e-12  6.95884232e-12
 -5.66883591e-11 -1.37449759e-11  3.14165930e-11 -1.56460616e-11
  1.26191219e-11  1.84601354e-11  1.66247040e-11  2.02430337e-11
  1.85940246e-11 -2.02599060e-11  1.68920620e-11 -4.37047044e-11
  8.23549006e-12  2.40167982e-10 -4.94651779e-11 -5.42201483e-11
  2.64697796e-23 -3.03742857e-11 -3.74767094e-11 -3.09055420e-11
 -3.77261190e-11 -1.34492009e-11 -3.17637355e-22 -3.11621415e-12
 -6.97464982e-12  2.93697126e-11 -5.27152820e-13 -9.30184492e-13
 -2.07113605e-11  4.35628604e-12  8.65423441e-12  5.61227654e-11
 -2.43762906e-11  1.31853253e-11 -4.63010398e-12 -3.69731664e-11
  1.02706641e-11 -1.04195234e-11 -5.27519563e-12 -2.21250431e-11
  2.93697126e-11 -2.39739029e-11 -9.50557616e-12 -2.43435791e-11
 -1.49991756e-11  2.57175092e-11  2.69176580e-11  3.10175538e-11
  2.48018590e-12 -4.84197343e-11 -2.17506330e-12  3.54635100e-11
 -7.73786103e-13  2.76883050e-12  9.30703928e-12 -6.89697905e-12
  3.66707011e-12 -6.00107628e-12  1.12532636e-11 -8.91925723e-12
 -3.73890917e-12 -3.36225164e-11 -1.26230439e-11  7.24835141e-11
 -2.24990447e-11 -1.57668488e-11 -2.11630702e-11 -5.99828362e-12
 -8.10592107e-12  1.98396175e-11  3.11945061e-11 -4.77554012e-13
  1.15049334e-11 -7.76920629e-12 -1.55541561e-11 -4.95748993e-12
 -4.84280206e-12 -2.27256383e-11  2.02299472e-11 -6.32827027e-12
 -1.86786692e-13 -8.66638642e-12 -6.93343305e-12  2.49949762e-12
 -3.70257281e-12 -1.64636182e-11  8.79825943e-12  3.46869478e-12
 -6.89624391e-12  2.37977853e-11  2.20289871e-11  1.48524701e-11
  4.49046707e-11  3.51287971e-11  6.24209784e-11  3.51882802e-11
 -1.02753364e-11 -2.79109580e-11 -8.26658963e-12  1.74813439e-11
 -6.21674007e-12  1.12765005e-11 -1.42652118e-11 -3.94940533e-12
  0.00000000e+00  6.93811644e-12  1.09463123e-12 -1.09463123e-12
 -2.22024598e-11 -5.69937149e-12 -1.31323641e-11 -2.27256383e-11
  7.82234549e-12  7.39911896e-12 -1.74093033e-11  1.21724159e-11
 -2.16223126e-12  1.93506687e-11  3.43888981e-12 -2.16641890e-11
 -1.72954805e-11  3.91069616e-12  3.15270866e-11 -1.92676719e-11
  1.89869225e-11  8.53336303e-12  1.63309181e-11  3.07955994e-12
 -2.29381390e-12 -2.77316890e-11  3.33962637e-14  2.69364318e-11
 -2.70121065e-12 -2.84043473e-11  3.60881198e-12 -3.30963175e-12
 -1.01982994e-11 -1.47506757e-11 -1.36936410e-10  2.41632963e-11
  3.32690215e-11  1.48567317e-11  3.63887536e-11  1.06352546e-10
 -3.62686006e-11 -3.89412289e-11 -2.57294383e-11 -3.36718847e-11
 -2.70743438e-12 -3.68253444e-12  6.38996888e-12 -1.86955479e-12
  2.78426664e-11 -2.37623929e-11 -4.53908842e-12 -3.16632170e-11
 -2.44306259e-12  1.54944750e-11  2.69189474e-11 -5.97877318e-12
  1.05328522e-11 -2.61469389e-11 -1.66131133e-11  8.39787560e-12
  9.90484172e-12  1.39244827e-11]
```

In [48]:

```python
predictions = lm.predict( X_test)
plt.scatter(y_test,predictions)
plt.xlabel('Y Test')
plt.ylabel('Predicted Y')
```

Out[48]:

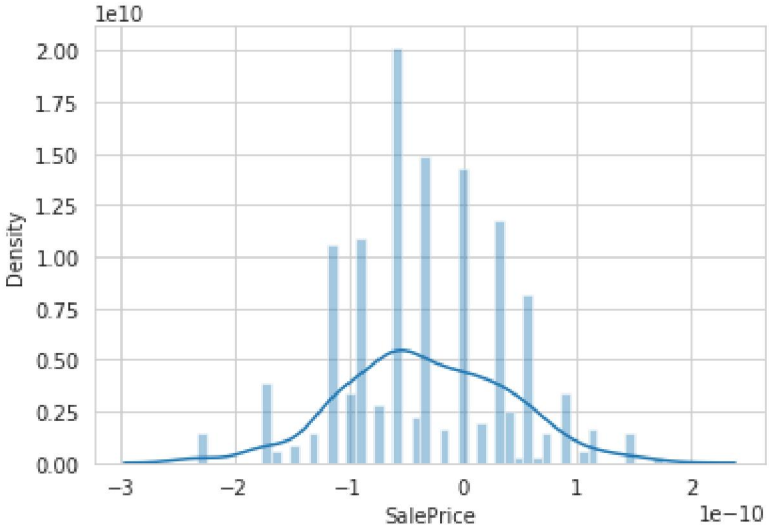
```
Text(0, 0.5, 'Predicted Y')
```



In [49]:

```
sns.distplot((y_test-predictions),bins=50);
```

```
/home/ec2-user/anaconda3/envs/tensorflow_p36/lib/python3.6/site-packages/sea
born/distributions.py:2557: FutureWarning: `distplot` is a deprecated functi
on and will be removed in a future version. Please adapt your code to use ei
ther `displot` (a figure-level function with similar flexibility) or `histpl
ot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```



In [50]:

```
coeffecients = pd.DataFrame(lm.coef_,X.columns)
coeffecients.columns = ['Coeffecient']
coeffecients
```

Out[50]:

	Coeffecient
Id	-4.985822e-15
MSSubClass	-8.062440e-13
LotArea	-1.110223e-15
OverallQual	4.794566e-12
OverallCond	-4.979778e-12
...	...
SaleCondition_AdjLand	-2.614694e-11
SaleCondition_Alloca	-1.661311e-11
SaleCondition_Family	8.397876e-12
SaleCondition_Normal	9.904842e-12
SaleCondition_Partial	1.392448e-11

274 rows × 1 columns

In [51]:

```
from sklearn import metrics

print('MAE:', metrics.mean_absolute_error(y_test, predictions))
print('MSE:', metrics.mean_squared_error(y_test, predictions))
print('RMSE:', np.sqrt(metrics.mean_squared_error(y_test, predictions)))
```

```
MAE: 6.319119352592181e-11
MSE: 6.176523640209654e-21
RMSE: 7.859086232005381e-11
```

In [ ]: