In [1]:

```
# This file provide a basic exploration of ames house price dataset
import numpy as np
import pandas as pd
```

In [2]:

```
df = pd.read_csv('./input/train.csv')
df.head()
```

Out[2]:

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

5 rows × 81 columns

In [3]:

```
df.describe()
```

```
/Users/swang/anaconda/lib/python2.7/site-packages/numpy/lib/functi
on_base.py:3834: RuntimeWarning: Invalid value encountered in perc
entile
  RuntimeWarning)
```

Out[3]:

	Id	MSSubClass	LotFrontage	LotArea	OverallQual	Ove
count	1460.000000	1460.000000	1201.000000	1460.000000	1460.000000	146
mean	730.500000	56.897260	70.049958	10516.828082	6.099315	5.57
std	421.610009	42.300571	24.284752	9981.264932	1.382997	1.11
min	1.000000	20.000000	21.000000	1300.000000	1.000000	1.00
25%	365.750000	20.000000	NaN	7553.500000	5.000000	5.00
50%	730.500000	50.000000	NaN	9478.500000	6.000000	5.00
75%	1095.250000	70.000000	NaN	11601.500000	7.000000	6.00
max	1460.000000	190.000000	313.000000	215245.000000	10.000000	9.00

8 rows × 38 columns

In [4]:

```
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
```

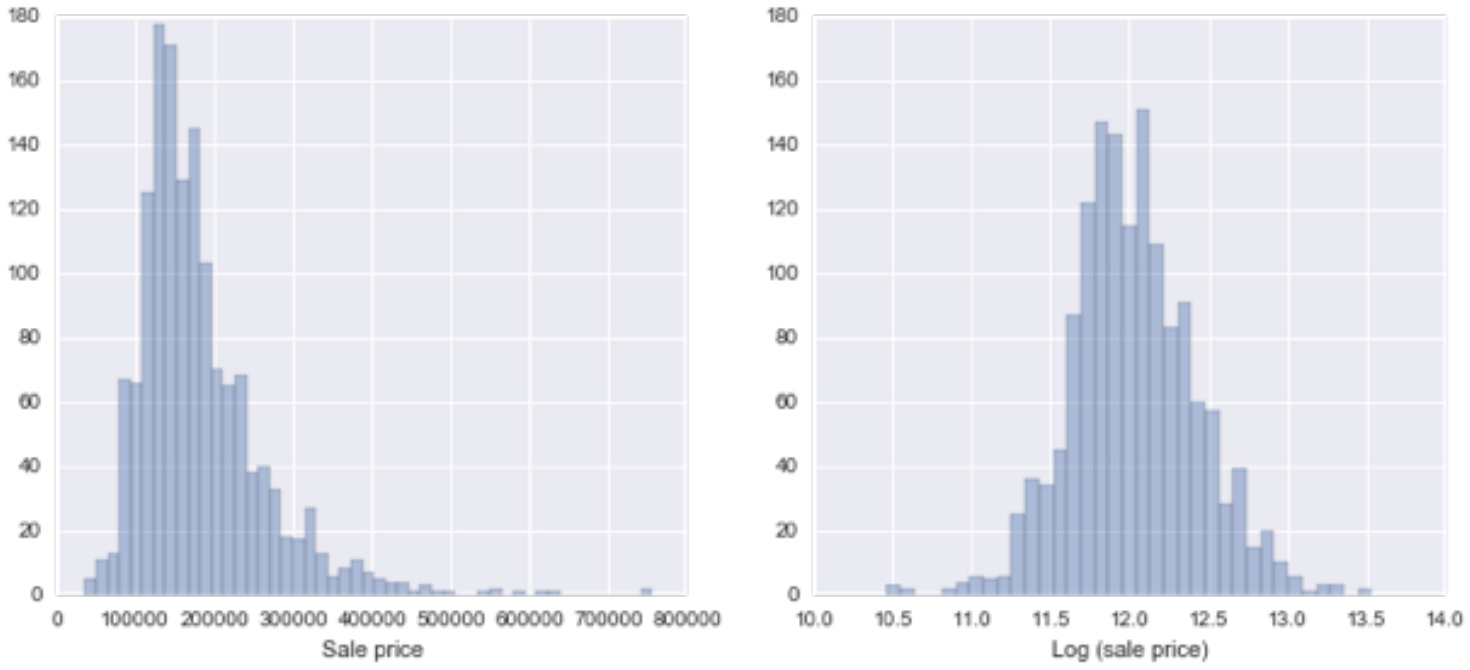
```
/Users/swang/anaconda/lib/python2.7/site-packages/matplotlib/font_
manager.py:273: UserWarning: Matplotlib is building the font cache
using fc-list. This may take a moment.
  warnings.warn('Matplotlib is building the font cache using fc-li
st. This may take a moment.')
```

In [5]:

```
# Set up the matplotlib figure
plt.figure(figsize=(12,5))
#f, axes = plt.subplots(1, 2, figsize=(12, 5), sharey=True)
plt.subplot(121)
sns.distplot(df['SalePrice'],kde=False)
plt.xlabel('Sale price')
plt.axis([0,800000,0,180])
plt.subplot(122)
sns.distplot(np.log(df['SalePrice']),kde=False)
plt.xlabel('Log (sale price)')
plt.axis([10,14,0,180])
```

Out[5]:

```
[10, 14, 0, 180]
```

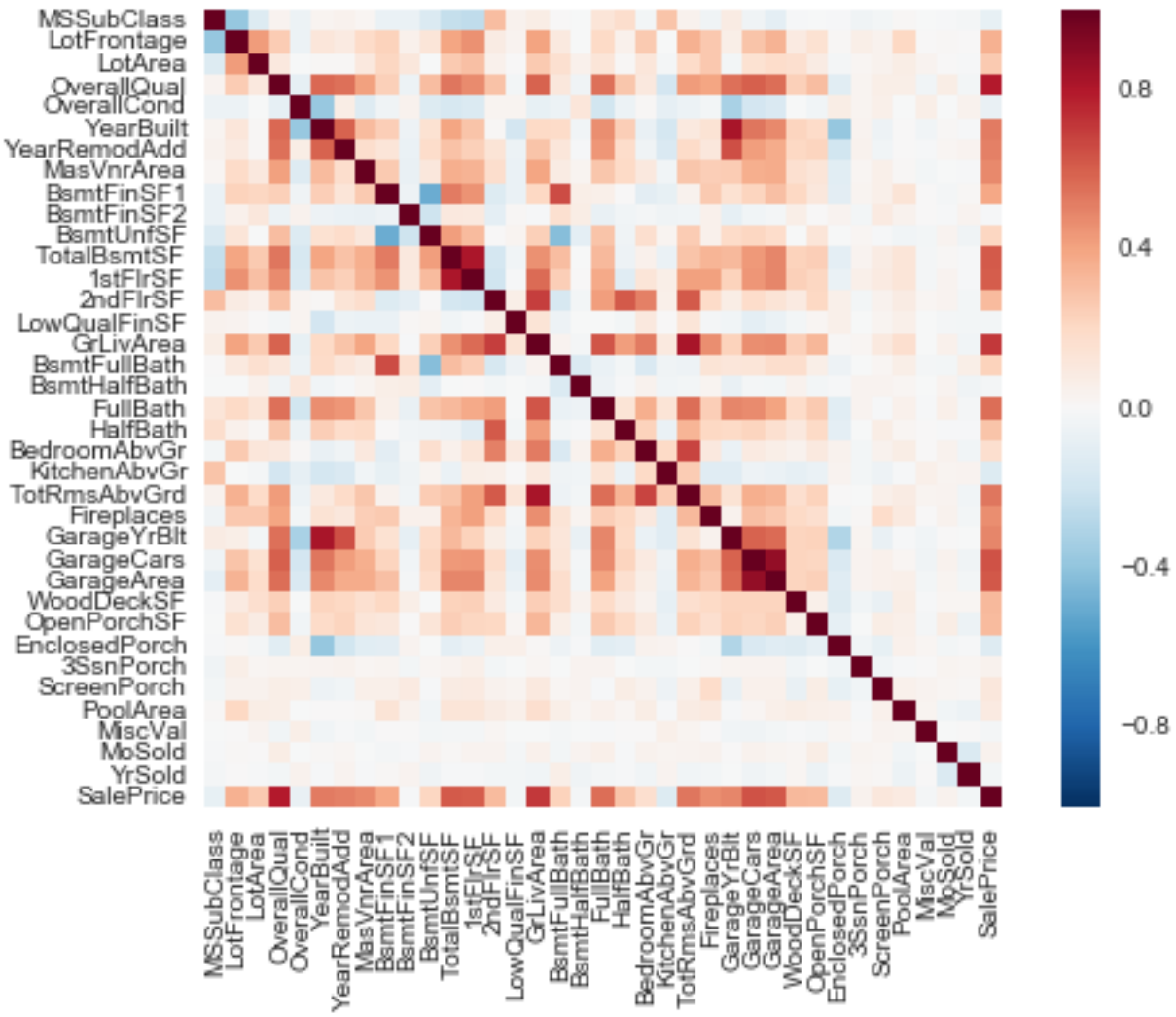


In [6]:

```
corr = df.select_dtypes(include = ['float64', 'int64']).iloc[:,1:].corr()
#fig = plt.figure()
sns.set(font_scale=1)
sns.heatmap(corr, vmax=1, square=True)
```

Out[6]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x11a142cd0>
```



In [7]:

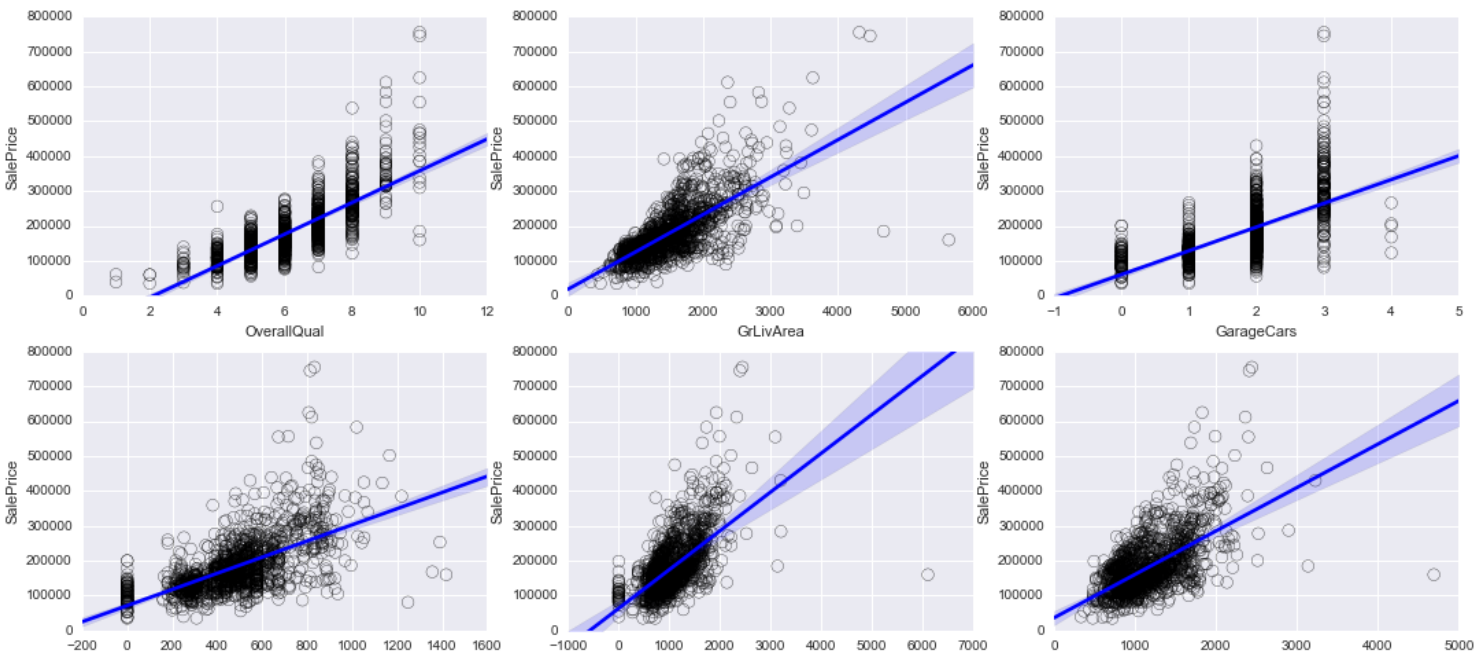
```
corr_list = corr['SalePrice'].sort_values(axis=0,ascending=False).iloc[1:]
corr_list
```

```
Out[7]:

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
YearRemodAdd     0.507101
GarageYrBlt      0.486362
MasVnrArea       0.477493
Fireplaces       0.466929
BsmtFinSF1       0.386420
LotFrontage      0.351799
WoodDeckSF       0.324413
2ndFlrSF         0.319334
OpenPorchSF      0.315856
HalfBath         0.284108
LotArea          0.263843
BsmtFullBath     0.227122
BsmtUnfSF        0.214479
BedroomAbvGr     0.168213
ScreenPorch      0.111447
PoolArea         0.092404
MoSold           0.046432
3SsnPorch        0.044584
BsmtFinSF2      -0.011378
BsmtHalfBath    -0.016844
MiscVal         -0.021190
LowQualFinSF    -0.025606
YrSold          -0.028923
OverallCond     -0.077856
MSSubClass      -0.084284
EnclosedPorch   -0.128578
KitchenAbvGr    -0.135907
Name: SalePrice, dtype: float64
```

In [8]:

```
plt.figure(figsize=(18,8))
for i in range(6):
    ii = '23'+str(i+1)
    plt.subplot(ii)
    feature = corr_list.index.values[i]
    plt.scatter(df[feature], df['SalePrice'], facecolors='none',edgecolors='k'
,s = 75)
    sns.regplot(x = feature, y = 'SalePrice', data = df,scatter=False, color =
'Blue')
    ax=plt.gca()
    ax.set_ylim([0,800000])
```



In [9]:

```
plt.figure(figsize = (12, 6))
sns.boxplot(x = 'Neighborhood', y = 'SalePrice',  data = df)
xt = plt.xticks(rotation=45)
```