cas

January 2, 2019

In [45]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
%matplotlib inline
```

In [46]:
```
train = pd.read_csv("train.csv")
test  = pd.read_csv("test.csv")
pd.options.display.max_columns=None
```

In [47]:
```
train.head()
```

Out[47]:

	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
0	1	60	RL	65.0	8450	Pave	NaN	Reg	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	2Story	7	5	2003	2003	Gable	CompShg	VinylSd	VinylSd	BrkFace
1	2	20	RL	80.0	9600	Pave	NaN	Reg	Lvl	AllPub	FR2	Gtl	Veenker	Feedr	Norm	1Fam	1Story	6	8	1976	1976	Gable	CompShg	MetalSd	MetalSd	None
2	3	60	RL	68.0	11250	Pave	NaN	IR1	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	2Story	7	5	2001	2002	Gable	CompShg	VinylSd	VinylSd	BrkFace
3	4	70	RL	60.0	9550	Pave	NaN	IR1	Lvl	AllPub	Corner	Gtl	Crawfor	Norm	Norm	1Fam	2Story	7	5	1915	1970	Gable	CompShg	Wd Sdng	Wd Shng	None
4	5	60	RL	84.0	14260	Pave	NaN	IR1	Lvl	AllPub	FR2	Gtl	NoRidge	Norm	Norm	1Fam	2Story	8	5	2000	2000	Gable	CompShg	VinylSd	VinylSd	BrkFace

```
4          2000     Gable  CompShg     VinylSd     VinylSd    BrkFace

   MasVnrArea ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure  \
0       196.0        Gd        TA      PConc       Gd       TA           No
1         0.0        TA        TA     CBlock       Gd       TA           Gd
2       162.0        Gd        TA      PConc       Gd       TA           Mn
3         0.0        TA        TA     BrkTil       TA       Gd           No
4       350.0        Gd        TA      PConc       Gd       TA           Av

  BsmtFinType1  BsmtFinSF1 BsmtFinType2  BsmtFinSF2  BsmtUnfSF  TotalBsmtSF  \
0          GLQ         706          Unf           0        150          856
1          ALQ         978          Unf           0        284         1262
2          GLQ         486          Unf           0        434          920
3          ALQ         216          Unf           0        540          756
4          GLQ         655          Unf           0        490         1145

  Heating HeatingQC CentralAir Electrical  1stFlrSF  2ndFlrSF  LowQualFinSF  \
0    GasA        Ex          Y      SBrkr       856       854             0
1    GasA        Ex          Y      SBrkr      1262         0             0
2    GasA        Ex          Y      SBrkr       920       866             0
3    GasA        Gd          Y      SBrkr       961       756             0
4    GasA        Ex          Y      SBrkr      1145      1053             0

   GrLivArea  BsmtFullBath  BsmtHalfBath  FullBath  HalfBath  BedroomAbvGr  \
0       1710             1             0         2         1             3
1       1262             0             1         2         0             3
2       1786             1             0         2         1             3
3       1717             1             0         1         0             3
4       2198             1             0         2         1             4

   KitchenAbvGr KitchenQual  TotRmsAbvGrd Functional  Fireplaces FireplaceQu  \
0             1          Gd             8        Typ           0         NaN
1             1          TA             6        Typ           1          TA
2             1          Gd             6        Typ           1          TA
3             1          Gd             7        Typ           1          Gd
4             1          Gd             9        Typ           1          TA

  GarageType  GarageYrBlt GarageFinish  GarageCars  GarageArea GarageQual  \
0     Attchd       2003.0          RFn           2         548         TA
1     Attchd       1976.0          RFn           2         460         TA
2     Attchd       2001.0          RFn           2         608         TA
3     Detchd       1998.0          Unf           3         642         TA
4     Attchd       2000.0          RFn           3         836         TA

  GarageCond PavedDrive  WoodDeckSF  OpenPorchSF  EnclosedPorch  3SsnPorch  \
0         TA          Y           0           61              0          0
1         TA          Y         298            0              0          0
2         TA          Y           0           42              0          0
```

```
3         TA          Y           0           35            272          0
4         TA          Y         192           84              0          0

   ScreenPorch  PoolArea PoolQC Fence MiscFeature  MiscVal  MoSold  YrSold  \
0            0         0    NaN   NaN         NaN        0       2    2008
1            0         0    NaN   NaN         NaN        0       5    2007
2            0         0    NaN   NaN         NaN        0       9    2008
3            0         0    NaN   NaN         NaN        0       2    2006
4            0         0    NaN   NaN         NaN        0      12    2008

  SaleType SaleCondition  SalePrice
0       WD        Normal     208500
1       WD        Normal     181500
2       WD        Normal     223500
3       WD       Abnorml     140000
4       WD        Normal     250000
```

```
In [48]: test.head()
```

```
Out[48]:      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 LotConfig LandSlope Neighborhood Condition1  \
0         Lvl    AllPub    Inside       Gtl        NAmes      Feedr
1         Lvl    AllPub    Corner       Gtl        NAmes       Norm
2         Lvl    AllPub    Inside       Gtl      Gilbert       Norm
3         Lvl    AllPub    Inside       Gtl      Gilbert       Norm
4         HLS    AllPub    Inside       Gtl      StoneBr       Norm

  Condition2 BldgType HouseStyle  OverallQual  OverallCond  YearBuilt  \
0       Norm     1Fam     1Story            5            6       1961
1       Norm     1Fam     1Story            6            6       1958
2       Norm     1Fam     2Story            5            5       1997
3       Norm     1Fam     2Story            6            6       1998
4       Norm   TwnhsE     1Story            8            5       1992

   YearRemodAdd RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType  \
0          1961     Gable  CompShg     VinylSd     VinylSd       None
1          1958       Hip  CompShg     Wd Sdng     Wd Sdng    BrkFace
2          1998     Gable  CompShg     VinylSd     VinylSd       None
3          1998     Gable  CompShg     VinylSd     VinylSd    BrkFace
4          1992     Gable  CompShg     HdBoard     HdBoard       None

   MasVnrArea ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure  \
```

0	0.0	TA	TA	CBlock	TA	TA	No
1	108.0	TA	TA	CBlock	TA	TA	No
2	0.0	TA	TA	PConc	Gd	TA	No
3	20.0	TA	TA	PConc	TA	TA	No
4	0.0	Gd	TA	PConc	Gd	TA	No

	BsmtFinType1	BsmtFinSF1	BsmtFinType2	BsmtFinSF2	BsmtUnfSF	TotalBsmtSF	\
0	Rec	468.0	LwQ	144.0	270.0	882.0	
1	ALQ	923.0	Unf	0.0	406.0	1329.0	
2	GLQ	791.0	Unf	0.0	137.0	928.0	
3	GLQ	602.0	Unf	0.0	324.0	926.0	
4	ALQ	263.0	Unf	0.0	1017.0	1280.0	

	Heating	HeatingQC	CentralAir	Electrical	1stFlrSF	2ndFlrSF	LowQualFinSF	\
0	GasA	TA	Y	SBrkr	896	0	0	
1	GasA	TA	Y	SBrkr	1329	0	0	
2	GasA	Gd	Y	SBrkr	928	701	0	
3	GasA	Ex	Y	SBrkr	926	678	0	
4	GasA	Ex	Y	SBrkr	1280	0	0	

	GrLivArea	BsmtFullBath	BsmtHalfBath	FullBath	HalfBath	BedroomAbvGr	\
0	896	0.0	0.0	1	0	2	
1	1329	0.0	0.0	1	1	3	
2	1629	0.0	0.0	2	1	3	
3	1604	0.0	0.0	2	1	3	
4	1280	0.0	0.0	2	0	2	

	KitchenAbvGr	KitchenQual	TotRmsAbvGrd	Functional	Fireplaces	FireplaceQu	\
0	1	TA	5	Typ	0	NaN	
1	1	Gd	6	Typ	0	NaN	
2	1	TA	6	Typ	1	TA	
3	1	Gd	7	Typ	1	Gd	
4	1	Gd	5	Typ	0	NaN	

	GarageType	GarageYrBlt	GarageFinish	GarageCars	GarageArea	GarageQual	\
0	Attchd	1961.0	Unf	1.0	730.0	TA	
1	Attchd	1958.0	Unf	1.0	312.0	TA	
2	Attchd	1997.0	Fin	2.0	482.0	TA	
3	Attchd	1998.0	Fin	2.0	470.0	TA	
4	Attchd	1992.0	RFn	2.0	506.0	TA	

	GarageCond	PavedDrive	WoodDeckSF	OpenPorchSF	EnclosedPorch	3SsnPorch	\
0	TA	Y	140	0	0	0	
1	TA	Y	393	36	0	0	
2	TA	Y	212	34	0	0	
3	TA	Y	360	36	0	0	
4	TA	Y	0	82	0	0	

```
   ScreenPorch  PoolArea PoolQC  Fence MiscFeature  MiscVal  MoSold  YrSold  \
0          120         0    NaN  MnPrv         NaN        0       6    2010
1            0         0    NaN    NaN        Gar2    12500       6    2010
2            0         0    NaN  MnPrv         NaN        0       3    2010
3            0         0    NaN    NaN         NaN        0       6    2010
4          144         0    NaN    NaN         NaN        0       1    2010

  SaleType SaleCondition
0       WD        Normal
1       WD        Normal
2       WD        Normal
3       WD        Normal
4       WD        Normal
```

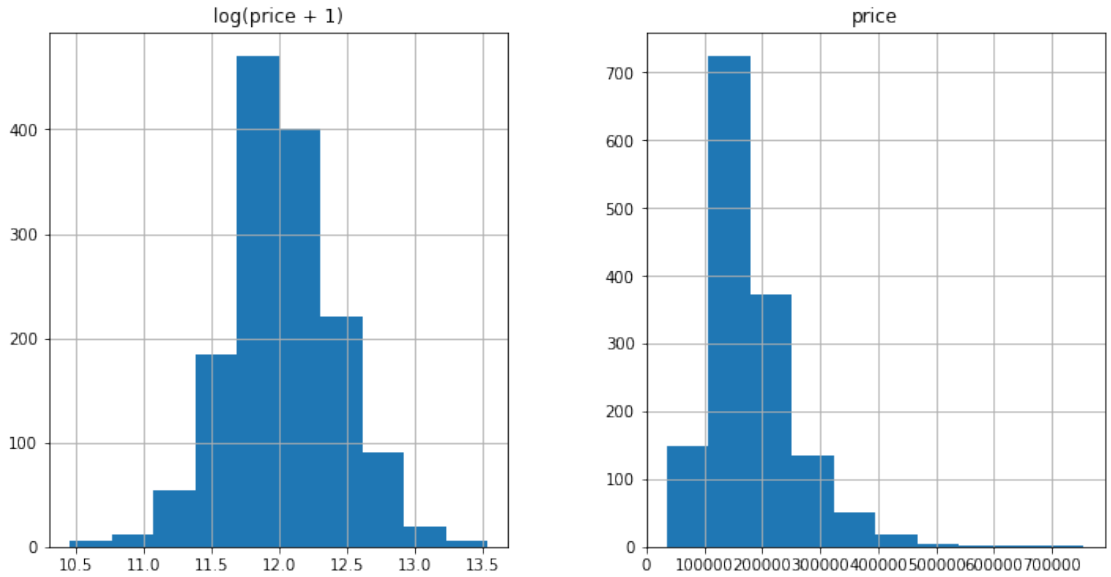
In [49]:
```
#all_data train+test data without target value
all_data = pd.concat((train.loc[:,"MSSubClass":"SaleCondition"],test.loc[:,"MSSubClass
```

In [50]:
```
all_data.shape
```

Out[50]:
```
(2919, 79)
```

In [56]:
```
#Feature engineering
plt.rcParams['figure.figsize'] = (12.0, 6.0)
prices = pd.DataFrame({"price":train["SalePrice"], "log(price + 1)":np.log1p(train["Sa
prices.hist()
```

Out[56]:
```
array([[<matplotlib.axes._subplots.AxesSubplot object at 0x7faf6ae0b9e8>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7faf6ae38fd0>]],
      dtype=object)
```



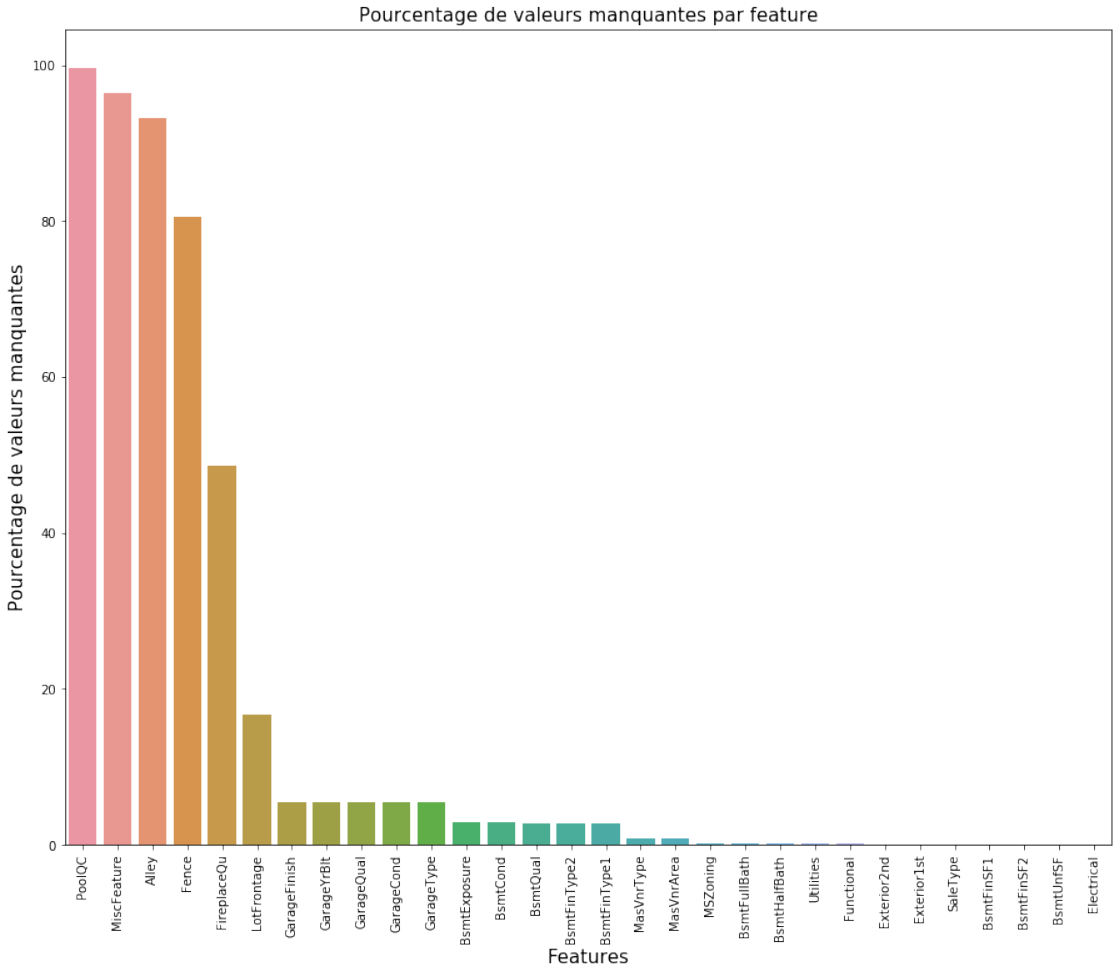
In [52]:
```
all_data_na = (all_data.isnull().sum()/ len(all_data)) *100
all_data_na = all_data_na.drop(all_data_na[all_data_na == 0].index).sort_values(ascend
missing_data = pd.DataFrame({'Missing Ratio' :all_data_na})
missing_data.head()
```

Out[52]:

	Missing Ratio
PoolQC	99.657417
MiscFeature	96.402878
Alley	93.216855
Fence	80.438506
FireplaceQu	48.646797

In [54]:
```
f, ax = plt.subplots(figsize=(15, 12))
plt.xticks(rotation='90')
sns.barplot(x=all_data_na.index, y=all_data_na)
plt.xlabel('Features', fontsize=15)
plt.ylabel('Pourcentage de valeurs manquantes', fontsize=15)
plt.title('Pourcentage de valeurs manquantes par feature', fontsize=15)
```

Out[54]: Text(0.5,1,'Pourcentage de valeurs manquantes par feature')



In [59]:
```
#matrice de correlation
corrmat = train.corr()
plt.subplots(figsize=(12,9))
sns.heatmap(corrmat, vmax=0.9, square=True)
```

Out[59]: <matplotlib.axes._subplots.AxesSubplot at 0x7faf6a7fb588>