In [1]:

```
import pandas as pd
import numpy as np
#pd.set_option('max_columns', 120)
#pd.set_option('max_colwidth', 5000)

import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
plt.rcParams['figure.figsize'] = (12,8)
```

In [2]:

```
loans = pd.read_csv('cleaned_loans_2007_test.csv',low_memory=False)
loans.head()
```

Out[2]:

	loan_amnt	installment	grade	emp_length	annual_inc	loan_status	dti	delinq_2yrs	inc
0	5000.0	162.87	2	10	24000.0	1	27.65	0.0	
1	2500.0	59.83	3	0	30000.0	0	1.00	0.0	
2	2400.0	84.33	3	10	12252.0	1	8.72	0.0	
3	10000.0	339.31	3	10	49200.0	1	20.00	0.0	
4	5000.0	156.46	1	3	36000.0	1	11.20	0.0	

5 rows × 39 columns

In [3]:

```
loans.shape
```

Out[3]:

```
(39177, 39)
```

In [4]:

```
loans['loan_status'].value_counts() / loans.shape[0]
```

Out[4]:

```
1    0.856191
0    0.143809
Name: loan_status, dtype: float64
```

In [5]:

```
loans.describe()
```

Out[5]:

	loan_amnt	installment	grade	emp_length	annual_inc	loan_status
count	39177.000000	39177.000000	39177.000000	39177.000000	3.917700e+04	39177.000000
mean	11143.689537	323.514635	2.565561	4.828471	6.891654e+04	0.856191
std	7398.202266	208.483501	1.383501	3.603729	6.400410e+04	0.350900
min	500.000000	15.690000	1.000000	0.000000	4.000000e+03	0.000000
25%	5425.000000	166.500000	1.000000	2.000000	4.020000e+04	1.000000
50%	10000.000000	279.160000	2.000000	4.000000	5.900000e+04	1.000000
75%	15000.000000	428.030000	3.000000	9.000000	8.200000e+04	1.000000
max	35000.000000	1305.190000	7.000000	10.000000	6.000000e+06	1.000000

8 rows × 39 columns

In [6]:

```
from sklearn.cross_validation import train_test_split
#from sklearn.model_selection import train_test_split
y = loans.pop('loan_status')
X = loans

X_train,X_test,y_train,y_test = train_test_split(X,y,stratify=y,test_size=0.25)
```

```
/Library/Frameworks/Python.framework/Versions/3.6/lib/python3.6/site
-packages/sklearn/cross_validation.py:41: DeprecationWarning: This m
odule was deprecated in version 0.18 in favor of the model_selection
module into which all the refactored classes and functions are moved
. Also note that the interface of the new CV iterators are different
from that of this module. This module will be removed in 0.20.
  "This module will be removed in 0.20.", DeprecationWarning)
```

In [7]:

```
print(X_train.shape)
```

```
(29382, 38)
```

In [8]:

```
print(X_test.shape)
```

```
(9795, 38)
```

In [9]:

```
y_train.value_counts() / y_train.shape[0]
```

Out[9]:

```
1    0.856204
0    0.143796
Name: loan_status, dtype: float64
```