Imports

In [21]:
```
from sqlalchemy import create_engine
import pymysql
import pandas as pd
import numpy as np
from matplotlib import pyplot as plt
import matplotlib.colors as mcolors
from matplotlib import cm
import pandas_profiling
import seaborn as sns
#sns.set(color_codes=True, font_scale=1.2)
# Import the two methods from heatmap library
from heatmap import heatmap, corrplot
```

Start with cleaned data from C2T2A notebook

In [22]:
```
df = pd.read_csv('creditOne_cleaned_formatted.csv', index_col=0)
df.rename(columns={'PAY_0':'PAY_1'}, inplace=True)
df.head(3)
```

Out[22]:

	ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_1	PAY_2	PAY_3	PAY_4	...	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_
0	1	20000	female	university	1	24	2	2	-1	-1	...	0	0	0	
1	2	120000	female	university	2	26	-1	2	0	0	...	3272	3455	3261	
2	3	90000	female	university	2	34	0	0	0	0	...	14331	14948	15549	

3 rows × 25 columns

In [23]:
```
df.dtypes.head()
```

Out[23]:
```
ID            int64
LIMIT_BAL     int64
SEX          object
EDUCATION    object
MARRIAGE      int64
dtype: object
```

Pandas profile may be more meaningful now that the data has been cleaned. May give me a good place to start my EDA

In [24]:
```
#rpt = pandas_profiling.ProfileReport(df)
#rpt.to_file(output_file='CreditOne_ProfileReport_CleanedData.html')
```

Use Dummy to split object columns

In [25]:
```
df = pd.get_dummies(df)
```

In [26]:
```
df.head()
```

Out[26]:

	ID	LIMIT_BAL	MARRIAGE	AGE	PAY_1	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6	...	PAY_AMT5	PAY_AMT6	SEX_female	SEX_male
0	1	20000	1	24	2	2	-1	-1	-2	-2	...	0	0	1	0
1	2	120000	2	26	-1	2	0	0	0	2	...	0	2000	1	0
2	3	90000	2	34	0	0	0	0	0	0	...	1000	5000	1	0
3	4	50000	1	37	0	0	0	0	0	0	...	1069	1000	1	0
4	5	50000	1	57	-1	0	-1	0	0	0	...	689	679	0	1

5 rows × 30 columns

In [27]:
```
df.dtypes
```

Out[27]:
```
ID                             int64
LIMIT_BAL                      int64
MARRIAGE                       int64
AGE                            int64
PAY_1                          int64
PAY_2                          int64
PAY_3                          int64
PAY_4                          int64
PAY_5                          int64
PAY_6                          int64
BILL_AMT1                      int64
BILL_AMT2                      int64
BILL_AMT3                      int64
BILL_AMT4                      int64
BILL_AMT5                      int64
BILL_AMT6                      int64
PAY_AMT1                       int64
PAY_AMT2                       int64
PAY_AMT3                       int64
PAY_AMT4                       int64
PAY_AMT5                       int64
PAY_AMT6                       int64
SEX_female                     uint8
SEX_male                       uint8
EDUCATION_graduate school      uint8
EDUCATION_high school          uint8
EDUCATION_other                uint8
EDUCATION_university           uint8
DEFAULT_default                uint8
DEFAULT_not default            uint8
dtype: object
```

EDA

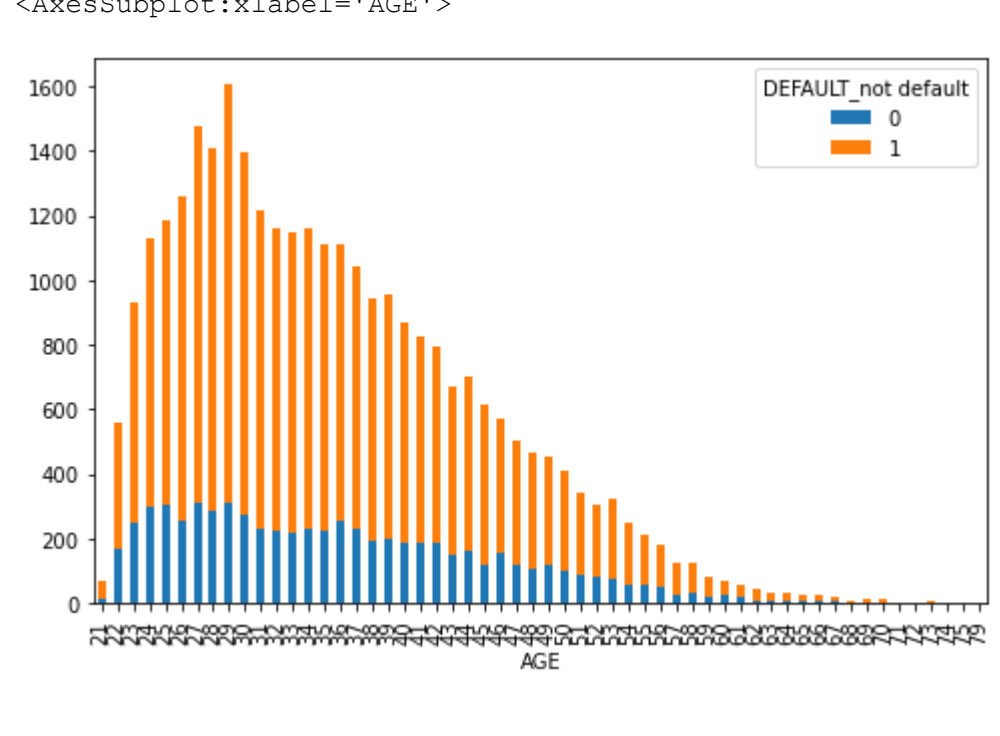
Here's some fixed vars I'll probably need for plotting:

In [28]:
```
PAY_XLABELS = ['No consumption', 'Paid in full', 'Revolving Credit',
               '1 month delay', '2 month delay', '3 month delay',
               '4 month delay', '5 month delay', '6 month delay', '7 month delay', '8 month delay']
```

will now regenerate the stacked column plot from my prior notebook, now that the columns have been split with dummy()

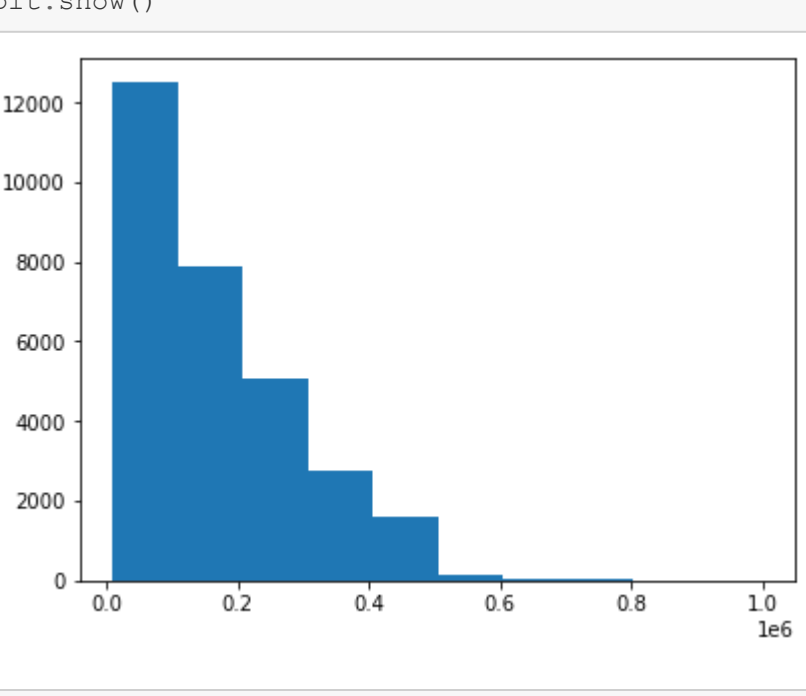
In [29]:
```
df.groupby('AGE')['DEFAULT_not default'].value_counts().unstack(level=1).plot.bar(stacked=True, figsize
=[8,5])
```

Out[29]:
```
<AxesSubplot:xlabel='AGE'>
```

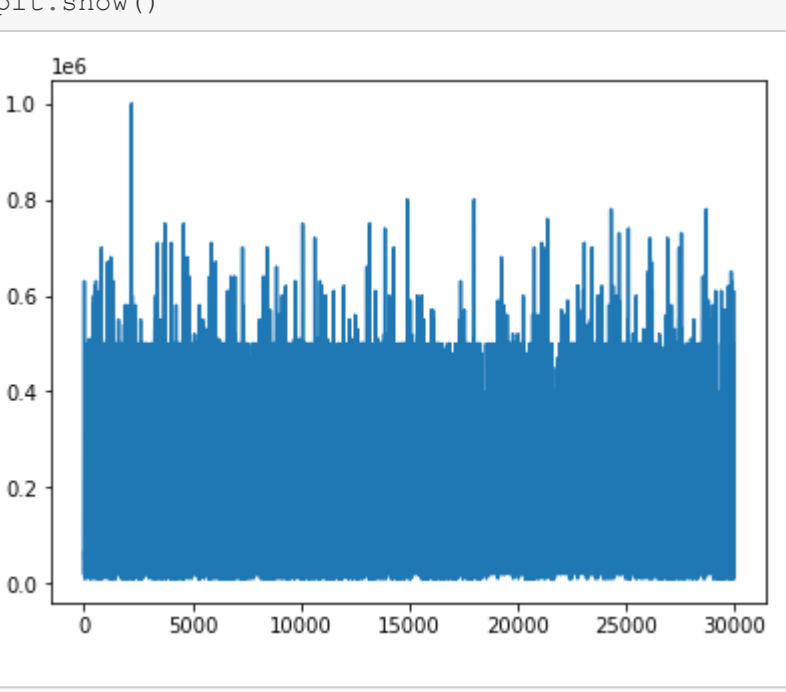


Plots from the Project Roadmap, for reference

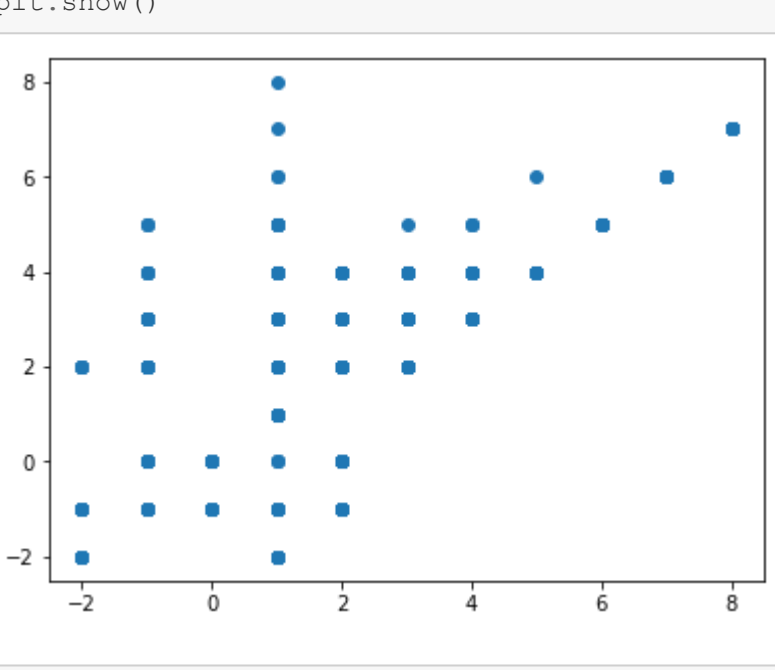
In [30]:
```
default_x=6.4
default_y=4.8
plt.rcParams['figure.figsize'] = (default_x,default_y)
plt.hist(df['LIMIT_BAL'], bins=10)
plt.show()
```



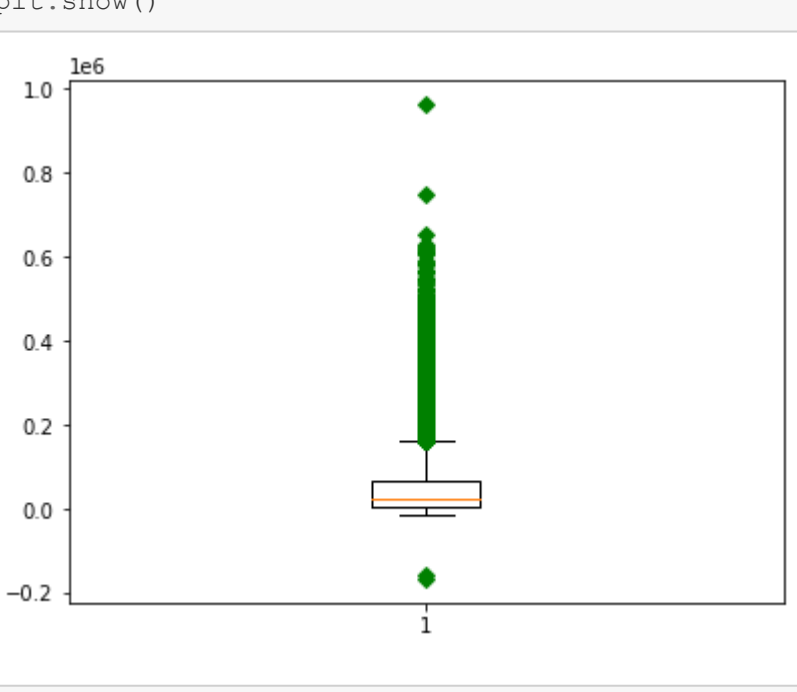
In [31]:
```
plt.plot(df['LIMIT_BAL'])
plt.show()
```



In [32]:
```
plt.scatter(df['PAY_1'],df['PAY_2'])
plt.show()
```



In [33]:
```
plt.boxplot(df['BILL_AMT1'],0,'gD')
plt.show()
```



In [34]:
```
corrMat = df.corr()
print(corrMat)
```

```
                                 ID  LIMIT_BAL  MARRIAGE       AGE     PAY_1  \
ID                         1.000000   0.026179 -0.029079  0.018678 -0.030575
LIMIT_BAL                  0.026179   1.000000 -0.108139  0.144713 -0.271214
MARRIAGE                  -0.029079  -0.108139  1.000000 -0.414170  0.019917
AGE                        0.018678   0.144713 -0.414170  1.000000 -0.039447
PAY_1                     -0.030575  -0.271214  0.019917 -0.039447  1.000000
PAY_2                     -0.011215  -0.296382  0.024199 -0.050148  0.672164
PAY_3                     -0.018494  -0.286123  0.032688 -0.053048  0.574245
PAY_4                     -0.002735  -0.267460  0.033122 -0.049722  0.538841
PAY_5                     -0.022199  -0.249411  0.035629 -0.053826  0.509426
PAY_6                     -0.020270  -0.235195  0.034345 -0.048773  0.474553
BILL_AMT1                  0.019389   0.285430 -0.023472  0.056239  0.187068
BILL_AMT2                  0.017982   0.278314 -0.021602  0.054283  0.189859
BILL_AMT3                  0.024354   0.283236 -0.024909  0.053710  0.179785
BILL_AMT4                  0.040351   0.293988 -0.023344  0.051353  0.179125
BILL_AMT5                  0.016705   0.295562 -0.025393  0.049345  0.180635
BILL_AMT6                  0.016730   0.290389 -0.021207  0.047613  0.176980
PAY_AMT1                   0.009742   0.195236 -0.005979  0.026147 -0.079269
PAY_AMT2                   0.008406   0.178408 -0.008093  0.021785 -0.070101
PAY_AMT3                   0.039151   0.210167 -0.003541  0.029247 -0.070561
PAY_AMT4                   0.007793   0.203242 -0.012659  0.021379 -0.064005
PAY_AMT5                   0.000652   0.217202 -0.001205  0.022850 -0.058190
PAY_AMT6                   0.003000   0.219595 -0.006641  0.019478 -0.058673
SEX_female                 0.018497   0.024755 -0.031389 -0.090874 -0.057643
SEX_male                  -0.018497  -0.024755  0.031389  0.090874  0.057643
EDUCATION_graduate school -0.025858   0.258777  0.142129 -0.100423 -0.142720
EDUCATION_high school      0.017475  -0.139686 -0.110845  0.231252  0.058902
EDUCATION_other            0.037165   0.013420 -0.008386  0.008982 -0.024937
EDUCATION_university       0.002569  -0.147530 -0.051797 -0.077626  0.099177
DEFAULT_default           -0.013952  -0.153520 -0.024339  0.013890  0.324794
DEFAULT_not default        0.013952   0.153520  0.024339 -0.013890 -0.324794

                              PAY_2     PAY_3     PAY_4     PAY_5     PAY_6  \
ID                        -0.011215 -0.018494 -0.002735 -0.022199 -0.020270
LIMIT_BAL                 -0.296382 -0.286123 -0.267460 -0.249411 -0.235195
MARRIAGE                   0.024199  0.032688  0.033122  0.035629  0.034345
AGE                       -0.050148 -0.053048 -0.049722 -0.053826 -0.048773
PAY_1                      0.672164  0.574245  0.538841  0.509426  0.474553
PAY_2                      1.000000  0.766552  0.662067  0.622780  0.575501
PAY_3                      0.766552  1.000000  0.777359  0.686775  0.632684
PAY_4                      0.662067  0.777359  1.000000  0.819835  0.716449
PAY_5                      0.622780  0.686775  0.819835  1.000000  0.816900
PAY_6                      0.575501  0.632684  0.716449  0.816900  1.000000
BILL_AMT1                  0.234887  0.208473  0.202812  0.206684  0.207373
BILL_AMT2                  0.235257  0.237295  0.225816  0.226913  0.226924
BILL_AMT3                  0.224146  0.227494  0.244983  0.243335  0.241181
BILL_AMT4                  0.222237  0.227202  0.245917  0.271915  0.266356
BILL_AMT5                  0.221348  0.225145  0.242902  0.269783  0.290894
BILL_AMT6                  0.219403  0.222327  0.239154  0.262509  0.285091
PAY_AMT1                  -0.080701  0.001295 -0.009362 -0.006089 -0.001496
PAY_AMT2                  -0.058990 -0.066793 -0.001944 -0.003191 -0.005223
PAY_AMT3                  -0.055901 -0.053311 -0.069235  0.009062  0.005834
PAY_AMT4                  -0.046858 -0.046067 -0.043461 -0.058299  0.019018
PAY_AMT5                  -0.037093 -0.035863 -0.033590 -0.033337 -0.046434
PAY_AMT6                  -0.036500 -0.035861 -0.026565 -0.023027 -0.025299
SEX_female                -0.070771 -0.066096 -0.060173 -0.055064 -0.044008
SEX_male                   0.070771  0.066096  0.060173  0.055064  0.044008
EDUCATION_graduate school -0.169215 -0.160209 -0.152402 -0.138709 -0.125123
EDUCATION_high school      0.064590  0.062461  0.058382  0.049577  0.041370
EDUCATION_other           -0.033118 -0.034435 -0.030998 -0.028822 -0.035733
EDUCATION_university       0.122364  0.115644  0.110340  0.103218  0.098013
DEFAULT_default            0.263551  0.235253  0.216614  0.204149  0.186866
DEFAULT_not default       -0.263551 -0.235253 -0.216614 -0.204149 -0.186866

                           ...  PAY_AMT5  PAY_AMT6  SEX_female  SEX_male  \
ID                         ...  0.000652  0.003000    0.018497 -0.018497
LIMIT_BAL                  ...  0.217202  0.219595    0.024755 -0.024755
MARRIAGE                   ... -0.001205 -0.006641   -0.031389  0.031389
AGE                        ...  0.022850  0.019478   -0.090874  0.090874
PAY_1                      ... -0.058190 -0.058673   -0.057643  0.057643
PAY_2                      ... -0.037093 -0.036500   -0.070771  0.070771
PAY_3                      ... -0.035863 -0.035861   -0.066096  0.066096
PAY_4                      ... -0.033590 -0.026565   -0.060173  0.060173
PAY_5                      ... -0.033337 -0.023027   -0.055064  0.055064
PAY_6                      ... -0.046434 -0.025299   -0.044008  0.044008
BILL_AMT1                  ...  0.167026  0.179341   -0.033642  0.033642
BILL_AMT2                  ...  0.157957  0.174256   -0.031183  0.031183
BILL_AMT3                  ...  0.179712  0.182326   -0.024563  0.024563
BILL_AMT4                  ...  0.160433  0.177637   -0.021880  0.021880
BILL_AMT5                  ...  0.141574  0.164184   -0.017005  0.017005
BILL_AMT6                  ...  0.307729  0.115494   -0.016733  0.016733
PAY_AMT1                   ...  0.148459  0.185735   -0.000242  0.000242
PAY_AMT2                   ...  0.180908  0.157634   -0.001391  0.001391
PAY_AMT3                   ...  0.159214  0.162740   -0.008597  0.008597
PAY_AMT4                   ...  0.151830  0.157834   -0.002229  0.002229
PAY_AMT5                   ...  1.000000  0.154896   -0.001667  0.001667
PAY_AMT6                   ...  0.154896  1.000000   -0.002766  0.002766
SEX_female                 ... -0.001667 -0.002766    1.000000 -1.000000
SEX_male                   ...  0.001667  0.002766   -1.000000  1.000000
EDUCATION_graduate school  ...  0.047226  0.050135   -0.022750  0.022750
EDUCATION_high school      ... -0.034768 -0.034613   -0.007650  0.007650
EDUCATION_other            ...  0.007395  0.016008    0.008498 -0.008498
EDUCATION_university       ... -0.021270 -0.026310    0.025353 -0.025353
DEFAULT_default            ... -0.055124 -0.053183   -0.039961  0.039961
DEFAULT_not default        ...  0.055124  0.053183    0.039961 -0.039961

                           EDUCATION_graduate school  EDUCATION_high school  \
ID                                         -0.025858               0.017475
LIMIT_BAL                                   0.258777              -0.139686
MARRIAGE                                    0.142129              -0.110845
AGE                                        -0.100423               0.231252
PAY_1                                      -0.142720               0.058902
PAY_2                                      -0.169215               0.064590
PAY_3                                      -0.160209               0.062461
PAY_4                                      -0.152402               0.058382
PAY_5                                      -0.138709               0.049577
PAY_6                                      -0.125123               0.041370
BILL_AMT1                                  -0.024045              -0.022006
BILL_AMT2                                  -0.020334              -0.022761
BILL_AMT3                                  -0.013112              -0.022807
BILL_AMT4                                  -0.003809              -0.031276
BILL_AMT5                                   0.000941              -0.031708
BILL_AMT6                                  -0.002525              -0.030981
PAY_AMT1                                    0.049811              -0.021310
PAY_AMT2                                    0.044400              -0.016674
PAY_AMT3                                    0.055982              -0.031726
PAY_AMT4                                    0.046119              -0.023554
PAY_AMT5                                    0.047226              -0.034768
PAY_AMT6                                    0.050135              -0.034613
SEX_female                                 -0.022750              -0.007650
SEX_male                                    0.022750               0.007650
EDUCATION_graduate school                   1.000000              -0.326917
EDUCATION_high school                      -0.326917               1.000000
EDUCATION_other                            -0.092951              -0.055736
EDUCATION_university                       -0.692075              -0.414989
DEFAULT_default                            -0.051328               0.032403
DEFAULT_not default                         0.051328              -0.032403

                           EDUCATION_other  EDUCATION_university  \
ID                                0.037165              0.002569
LIMIT_BAL                         0.013420             -0.147530
MARRIAGE                         -0.008386             -0.051797
AGE                               0.008982             -0.077626
PAY_1                            -0.024937              0.099177
PAY_2                            -0.033118              0.122364
PAY_3                            -0.034435              0.115644
PAY_4                            -0.030998              0.110340
PAY_5                            -0.028822              0.103218
PAY_6                            -0.035733              0.098013
BILL_AMT1                         0.036364              0.030323
BILL_AMT2                         0.030778              0.028716
BILL_AMT3                         0.028952              0.022288
BILL_AMT4                         0.020954              0.021648
BILL_AMT5                         0.011811              0.019690
BILL_AMT6                         0.003417              0.024555
PAY_AMT1                          0.004445             -0.032998
PAY_AMT2                          0.012004             -0.033132
PAY_AMT3                          0.022272             -0.035608
PAY_AMT4                          0.001095             -0.026965
PAY_AMT5                          0.007395             -0.021270
PAY_AMT6                          0.016008             -0.026310
SEX_female                        0.008498              0.025353
SEX_male                         -0.008498             -0.025353
EDUCATION_graduate school        -0.092951             -0.692075
EDUCATION_high school            -0.055736             -0.414989
EDUCATION_other                   1.000000             -0.117992
EDUCATION_university             -0.117992              1.000000
DEFAULT_default                  -0.045703              0.036467
DEFAULT_not default               0.045703             -0.036467

                           DEFAULT_default  DEFAULT_not default
ID                               -0.013952             0.013952
LIMIT_BAL                        -0.153520             0.153520
MARRIAGE                         -0.024339             0.024339
AGE                               0.013890            -0.013890
PAY_1                             0.324794            -0.324794
PAY_2                             0.263551            -0.263551
PAY_3                             0.235253            -0.235253
PAY_4                             0.216614            -0.216614
PAY_5                             0.204149            -0.204149
PAY_6                             0.186866            -0.186866
BILL_AMT1                        -0.019644             0.019644
BILL_AMT2                        -0.014193             0.014193
BILL_AMT3                        -0.014076             0.014076
BILL_AMT4                        -0.010156             0.010156
BILL_AMT5                        -0.006760             0.006760
BILL_AMT6                        -0.005372             0.005372
PAY_AMT1                         -0.072929             0.072929
PAY_AMT2                         -0.058579             0.058579
PAY_AMT3                         -0.056250             0.056250
PAY_AMT4                         -0.056827             0.056827
PAY_AMT5                         -0.055124             0.055124
PAY_AMT6                         -0.053183             0.053183
SEX_female                       -0.039961             0.039961
SEX_male                          0.039961            -0.039961
EDUCATION_graduate school        -0.051328             0.051328
EDUCATION_high school             0.032403            -0.032403
EDUCATION_other                  -0.045703             0.045703
EDUCATION_university              0.036467            -0.036467
DEFAULT_default                   1.000000            -1.000000
DEFAULT_not default              -1.000000             1.000000

[30 rows x 30 columns]
```

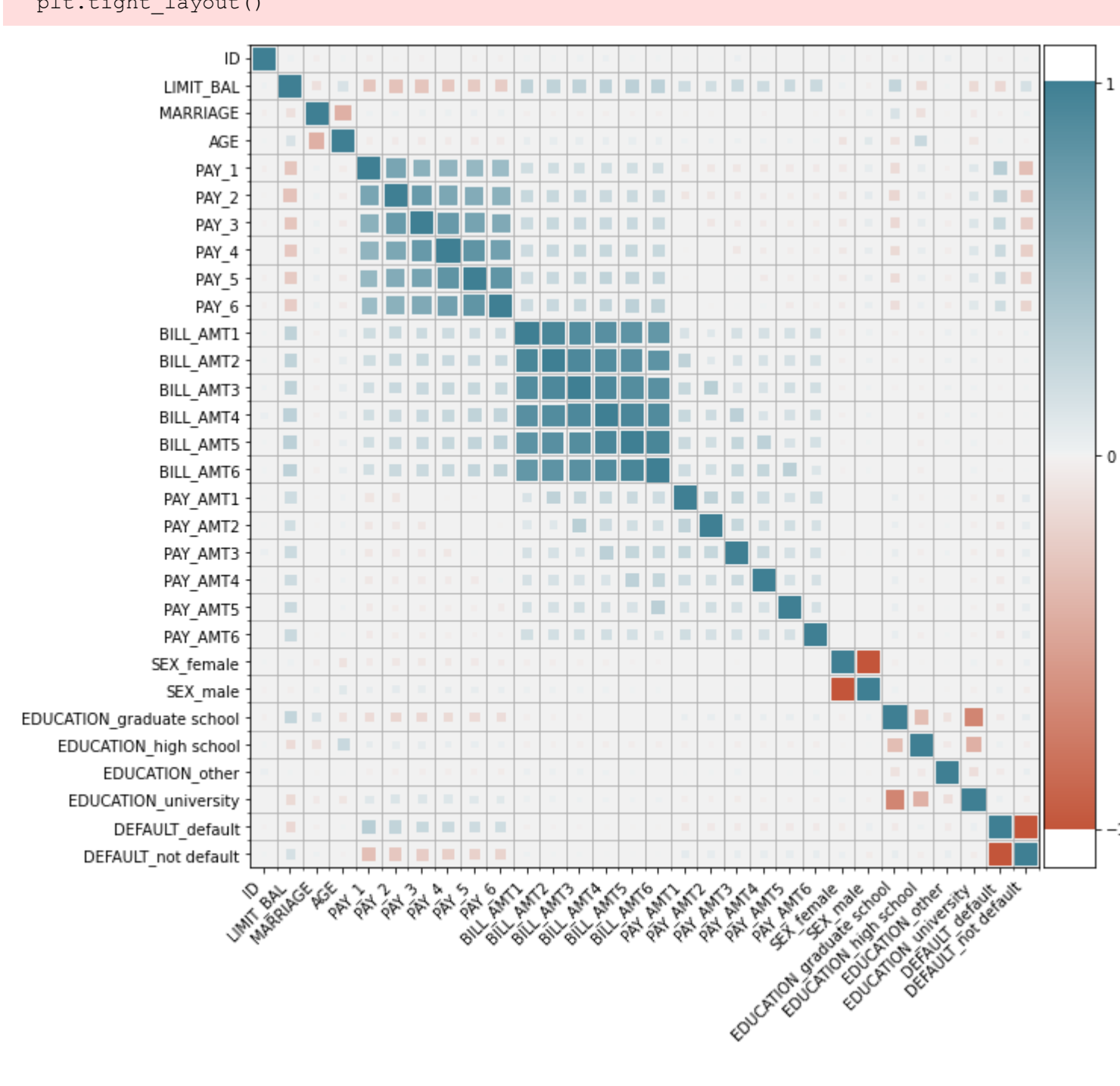
In [35]:
```
covMat = df.cov()
print(covMat)
```

```
                                     ID     LIMIT_BAL     MARRIAGE  \
ID                         7.500250e+07  2.941664e+07  -131.452515
LIMIT_BAL                  2.941664e+07  1.683446e+10 -7323.669658
MARRIAGE                  -1.314525e+02 -7.323670e+03     0.272452
AGE                        1.491057e+03  1.730767e+05    -1.992764
PAY_1                     -2.975726e+02 -3.954593e+04     0.011683
PAY_2                     -1.162770e+02 -4.603765e+04     0.015122
PAY_3                     -1.916922e+02 -4.443225e+04     0.020421
PAY_4                     -2.769132e+01 -4.057181e+04     0.020213
PAY_5                     -2.178600e+02 -3.667056e+04     0.021074
PAY_6                     -2.018774e+02 -3.509308e+04     0.020616
BILL_AMT1                  1.236447e+07  2.727020e+09  -902.154685
BILL_AMT2                  1.108392e+07  2.570130e+09  -802.517866
BILL_AMT3                  1.462688e+07  2.548533e+09  -901.679085
BILL_AMT4                  2.248126e+07  2.453926e+09  -783.881599
BILL_AMT5                  8.795803e+06  2.331481e+09  -805.840875
BILL_AMT6                  8.628506e+06  2.243837e+09  -659.223347
PAY_AMT1                   1.397501e+06  4.195711e+08   -51.691615
PAY_AMT2                   1.677383e+06  5.333504e+08   -97.327974
PAY_AMT3                   5.969803e+06  4.801180e+08   -32.546082
PAY_AMT4                   1.057334e+06  4.131202e+08  -103.518204
PAY_AMT5                   8.629502e+04  4.305657e+08    -9.607709
PAY_AMT6                   4.618460e+05  5.065153e+08   -61.623271
SEX_female                 7.835638e+01  1.571051e+03    -0.008014
SEX_male                  -7.835638e+01 -1.571051e+03     0.008014
EDUCATION_graduate school -1.070136e+02  1.604448e+04     0.035451
EDUCATION_high school      5.602518e+01 -6.709315e+03    -0.021418
EDUCATION_other            3.988670e+01  2.157851e+02    -0.000542
EDUCATION_university       1.110174e+01 -9.550953e+03    -0.013490
DEFAULT_default           -5.015171e+01 -8.267552e+03    -0.005273
DEFAULT_not default        5.015171e+01  8.267552e+03     0.005273

                                    AGE         PAY_1         PAY_2  \
ID                          1491.057019   -297.572636   -116.276959
LIMIT_BAL                 173076.722569 -39545.930009 -46037.648360
MARRIAGE                      -1.992764      0.011683      0.015122
AGE                           84.969755     -0.408639     -0.553408
PAY_1                         -0.408639      1.262930      0.904330
PAY_2                         -0.553408      0.904330      1.433254
PAY_3                         -0.585263      0.772384      1.098371
PAY_4                         -0.535851      0.707972      0.926680
PAY_5                         -0.562245      0.648743      0.844886
PAY_6                         -0.517022      0.613292      0.792320
BILL_AMT1                  38172.933546  15480.304170  20706.614217
BILL_AMT2                  35613.657962  15185.916919  20045.829482
BILL_AMT3                  34334.251320  14011.556537  18609.510991
BILL_AMT4                  30453.108180  12950.248389  17116.298983
BILL_AMT5                  27654.067800  12341.668685  16110.952468
BILL_AMT6                  26137.648547  11844.759724  15642.875812
PAY_AMT1                    3992.041735  -1475.495089  -1600.240756
PAY_AMT2                    4626.861549  -1815.138407  -1627.192336
PAY_AMT3                    4746.824393  -1396.168258  -1178.331282
PAY_AMT4                    3087.324192  -1126.847945   -878.843879
PAY_AMT5                    3218.052172   -999.107730   -678.468530
PAY_AMT6                    3191.903901  -1172.193614   -776.835035
SEX_female                    -0.409726     -0.031685     -0.041442
SEX_male                       0.409726      0.031685      0.041442
EDUCATION_graduate school     -0.442349     -0.076644     -0.096806
EDUCATION_high school          0.789120      0.024505      0.028625
EDUCATION_other                0.010260     -0.003473     -0.004913
EDUCATION_university          -0.357031      0.055612      0.073094
DEFAULT_default                0.053143      0.151499      0.130960
DEFAULT_not default           -0.053143     -0.151499     -0.130960

                                  PAY_3         PAY_4         PAY_5  \
ID                          -191.692156    -27.691323   -217.860029
LIMIT_BAL                 -44432.253315 -40571.811859 -36670.562325
MARRIAGE                       0.020421      0.020213      0.021074
AGE                           -0.585263     -0.535851     -0.562245
PAY_1                          0.772384      0.707972      0.648743
PAY_2                          1.098371      0.926680      0.844886
PAY_3                          1.432492      1.087761      0.931455
PAY_4                          1.087761      1.366885      1.086161
PAY_5                          0.931455      1.086161      1.284114
PAY_6                          0.870815      0.963263      1.064545
BILL_AMT1                  18373.210469  17460.198259  17246.377531
BILL_AMT2                  20214.071495  18790.627741  18301.285286
BILL_AMT3                  18882.491544  19862.999426  19122.663330
BILL_AMT4                  17494.100555  18496.423186  19822.925512
BILL_AMT5                  16382.947539  17265.551898  18586.590324
BILL_AMT6                  15847.089648  16651.586314  17715.690075
PAY_AMT1                      25.668468   -181.295613   -114.281714
PAY_AMT2                   -1841.952825    -52.358166    -83.324487
PAY_AMT3                   -1123.428782  -1425.205189    180.812142
PAY_AMT4                    -863.762183   -796.035739  -1034.961970
PAY_AMT5                    -655.796002   -599.991629   -577.161017
PAY_AMT6                    -763.026041   -552.137338   -463.892613
SEX_female                    -0.038694     -0.034411     -0.030521
SEX_male                       0.038694      0.034411      0.030521
EDUCATION_graduate school     -0.091629     -0.085144     -0.075112
EDUCATION_high school          0.027674      0.025268      0.020798
EDUCATION_other               -0.005107     -0.004491     -0.004047
EDUCATION_university           0.069062      0.064367      0.058361
DEFAULT_default                0.116867      0.105115      0.096020
DEFAULT_not default           -0.116867     -0.105115     -0.096020

                                  PAY_6  ...      PAY_AMT5      PAY_AMT6  \
ID                          -201.877379  ...  8.629502e+04  4.618460e+05
LIMIT_BAL                 -35093.083441  ...  4.305657e+08  5.065153e+08
MARRIAGE                       0.020616  ... -9.607709e+00 -6.162327e+01
AGE                           -0.517022  ...  3.218052e+03  3.191904e+03
PAY_1                          0.613292  ... -9.991077e+02 -1.172194e+03
PAY_2                          0.792320  ... -6.784685e+02 -7.768350e+02
PAY_3                          0.870815  ... -6.557960e+02 -7.630260e+02
PAY_4                          0.963263  ... -5.999916e+02 -5.521373e+02
PAY_5                          1.064545  ... -5.771610e+02 -4.638926e+02
PAY_6                          1.322472  ... -8.158327e+02 -5.172163e+02
BILL_AMT1                  17560.424872  ...  1.879091e+08  2.347681e+08
BILL_AMT2                  18573.527165  ...  1.717652e+08  2.204845e+08
BILL_AMT3                  19234.422476  ...  1.904126e+08  2.247817e+08
BILL_AMT4                  19705.551629  ...  1.576892e+08  2.031590e+08
BILL_AMT5                  20338.120325  ...  1.315051e+08  1.774537e+08
BILL_AMT6                  19524.880348  ...  2.799982e+08  1.222761e+08
PAY_AMT1                     -28.500666  ...  3.756893e+07  5.469033e+07
PAY_AMT2                    -138.399452  ...  6.368414e+07  6.456816e+07
PAY_AMT3                     118.121022  ...  4.282921e+07  5.093879e+07
PAY_AMT4                     342.623730  ...  3.634098e+07  4.395747e+07
PAY_AMT5                    -815.832688  ...  2.334266e+08  4.207110e+07
PAY_AMT6                    -517.216277  ...  4.207110e+07  3.160383e+08
SEX_female                    -0.024754  ... -1.245881e+01 -2.405188e+01
SEX_male                       0.024754  ...  1.245881e+01  2.405188e+01
EDUCATION_graduate school     -0.068759  ...  3.447913e+02  4.259025e+02
EDUCATION_high school          0.017612  ... -1.966421e+02 -2.277882e+02
EDUCATION_other               -0.005092  ...  1.400045e+01  3.526644e+01
EDUCATION_university           0.056240  ... -1.621496e+02 -2.333807e+02
DEFAULT_default                0.089194  ... -3.495625e+02 -3.924264e+02
DEFAULT_not default           -0.089194  ...  3.495625e+02  3.924264e+02

                            SEX_female     SEX_male  \
ID                           78.356379   -78.356379
LIMIT_BAL                  1571.050630 -1571.050630
MARRIAGE                     -0.008014     0.008014
AGE                          -0.409726     0.409726
PAY_1                        -0.031685     0.031685
PAY_2                        -0.041442     0.041442
PAY_3                        -0.038694     0.038694
PAY_4                        -0.034411     0.034411
PAY_5                        -0.030521     0.030521
PAY_6                        -0.024754     0.024754
BILL_AMT1                 -1211.694332  1211.694332
BILL_AMT2                 -1085.595467  1085.595467
BILL_AMT3                  -833.207432   833.207432
BILL_AMT4                  -688.489572   688.489572
BILL_AMT5                  -505.694333   505.694333
BILL_AMT6                  -487.430160   487.430160
PAY_AMT1                     -1.964266     1.964266
PAY_AMT2                    -15.675500    15.675500
PAY_AMT3                    -74.034812    74.034812
PAY_AMT4                    -17.080110    17.080110
PAY_AMT5                    -12.458809    12.458809
PAY_AMT6                    -24.051885    24.051885
SEX_female                    0.239247    -0.239247
SEX_male                     -0.239247     0.239247
EDUCATION_graduate school    -0.005317     0.005317
EDUCATION_high school        -0.001385     0.001385
EDUCATION_other               0.000515    -0.000515
EDUCATION_university          0.006188    -0.006188
DEFAULT_default              -0.008113     0.008113
DEFAULT_not default           0.008113    -0.008113

                           EDUCATION_graduate school  EDUCATION_high school  \
ID                                       -107.013617              56.025184
LIMIT_BAL                               16044.482969           -6709.314796
MARRIAGE                                    0.035451              -0.021418
AGE                                        -0.442349               0.789120
PAY_1                                      -0.076644               0.024505
PAY_2                                      -0.096806               0.028625
PAY_3                                      -0.091629               0.027674
PAY_4                                      -0.085144               0.025268
PAY_5                                      -0.075112               0.020798
PAY_6                                      -0.068759               0.017612
BILL_AMT1                                -846.084922            -599.858730
BILL_AMT2                                -691.567207            -599.693276
BILL_AMT3                                -434.516602            -585.509355
BILL_AMT4                                -117.103564            -744.846531
BILL_AMT5                                  27.342070            -713.633173
BILL_AMT6                                 -71.869023            -683.020797
PAY_AMT1                                  394.252555            -130.662666
PAY_AMT2                                  488.852307            -142.222305
PAY_AMT3                                  471.014411            -206.787191
PAY_AMT4                                  345.254954            -136.601819
PAY_AMT5                                  344.791256            -196.642088
PAY_AMT6                                  425.902474            -227.788197
SEX_female                                 -0.005317              -0.001385
SEX_male                                    0.005317               0.001385
EDUCATION_graduate school                   0.228350              -0.057831
EDUCATION_high school                      -0.057831               0.137041
EDUCATION_other                            -0.005504              -0.002557
EDUCATION_university                       -0.165014              -0.076653
DEFAULT_default                            -0.010180               0.004979
DEFAULT_not default                         0.010180              -0.004979

                           EDUCATION_other  EDUCATION_university  \
ID                               39.886696             11.101737
LIMIT_BAL                       215.785093          -9550.953266
MARRIAGE                         -0.000542             -0.013490
AGE                               0.010260             -0.357031
PAY_1                            -0.003473              0.055612
PAY_2                            -0.004913              0.073094
PAY_3                            -0.005107              0.069062
PAY_4                            -0.004491              0.064367
PAY_5                            -0.004047              0.058361
PAY_6                            -0.005092              0.056240
BILL_AMT1                       331.829532           1114.114120
BILL_AMT2                       271.462376           1019.798107
BILL_AMT3                       248.814746            771.211212
BILL_AMT4                       167.053398            694.896696
BILL_AMT5                        88.983544            597.307558
BILL_AMT6                        25.219712            729.670109
PAY_AMT1                          9.123682           -272.713571
PAY_AMT2                         34.274825           -380.904827
PAY_AMT3                         48.596788           -312.824009
PAY_AMT4                          2.126372           -210.779507
PAY_AMT5                         14.000453           -162.149622
PAY_AMT6                         35.266436           -233.380713
SEX_female                        0.000515              0.006188
SEX_male                         -0.000515             -0.006188
EDUCATION_graduate school        -0.005504             -0.165014
EDUCATION_high school            -0.002557             -0.076653
EDUCATION_other                   0.015357             -0.007296
EDUCATION_university             -0.007296              0.248963
DEFAULT_default                  -0.002351              0.007552
DEFAULT_not default               0.002351             -0.007552

                           DEFAULT_default  DEFAULT_not default
ID                              -50.151705            50.151705
LIMIT_BAL                     -8267.551759          8267.551759
MARRIAGE                         -0.005273             0.005273
AGE                               0.053143            -0.053143
PAY_1                             0.151499            -0.151499
PAY_2                             0.130960            -0.130960
PAY_3                             0.116867            -0.116867
PAY_4                             0.105115            -0.105115
PAY_5                             0.096020            -0.096020
PAY_6                             0.089194            -0.089194
BILL_AMT1                      -600.394108           600.394108
BILL_AMT2                      -419.289137           419.289137
BILL_AMT3                      -405.153680           405.153680
BILL_AMT4                      -271.199885           271.199885
BILL_AMT5                      -170.597447           170.597447
BILL_AMT6                      -132.796294           132.796294
PAY_AMT1                       -501.374552           501.374552
PAY_AMT2                       -560.210740           560.210740
PAY_AMT3                       -411.076284           411.076284
PAY_AMT4                       -369.515887           369.515887
PAY_AMT5                       -349.562530           349.562530
PAY_AMT6                       -392.426415           392.426415
SEX_female                       -0.008113             0.008113
SEX_male                          0.008113            -0.008113
EDUCATION_graduate school        -0.010180             0.010180
EDUCATION_high school             0.004979            -0.004979
EDUCATION_other                  -0.002351             0.002351
EDUCATION_university              0.007552            -0.007552
DEFAULT_default                   0.172276            -0.172276
DEFAULT_not default              -0.172276             0.172276

[30 rows x 30 columns]
```

Highest covarience is between Bill Amts and LIMIT_BAL. Makes sense... higher balances spend more...

In [36]:
```
plt.figure(figsize=(10, 10))
corrplot(df.corr(), size_scale=200);
plt.savefig('heatmap.png')
plt.tight_layout()
plt.show()
```

```
<ipython-input-36-04552ca69cd2>:4: UserWarning: This figure includes Axes that are not compatible wit
h tight_layout, so results might be incorrect.
  plt.tight_layout()
```



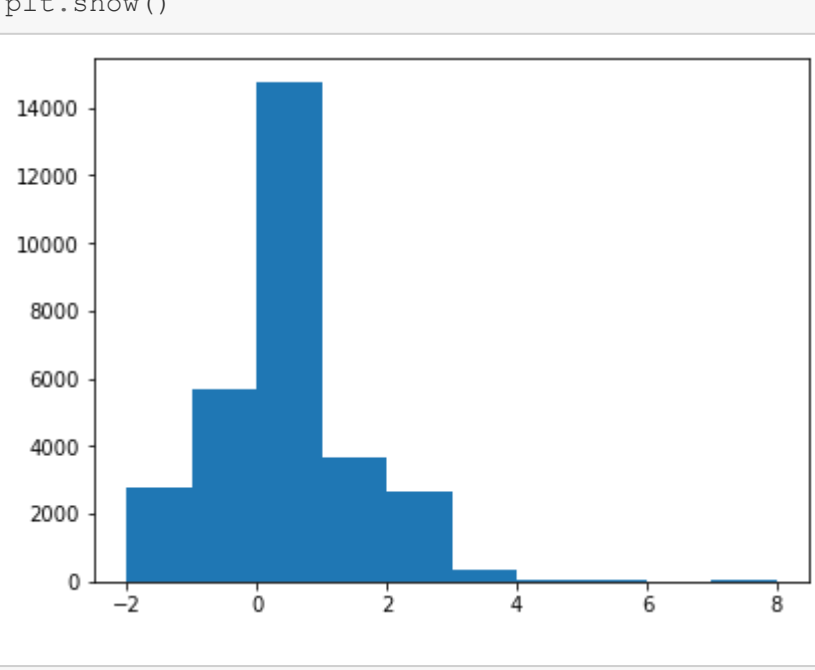
Initial thoughts

default does have some inverse relation to LIMIT_BAL. Also to Pay1-6 in decreasing order. remember - Pay1 is the most recent. Plot ideas:

- BILL AMT as box plots next to eachother - high correlation... are billing amounts that constant? look at % monthly change per user maybe? How would that link to defaults? (wild spending habits vs. stable)
- PAY - same story.. look into how constant customers are. Maybe even look at the sum change in value for a customer over their history - how constant are they. Corr makes me think it may be pretty constant
- LIMIT_BAL vs. DEFAULT ... somehow
- PAY_1 vs. DEFAULT ... stack by PAY_1 values
- Consider binning bill amounts or pay amounts

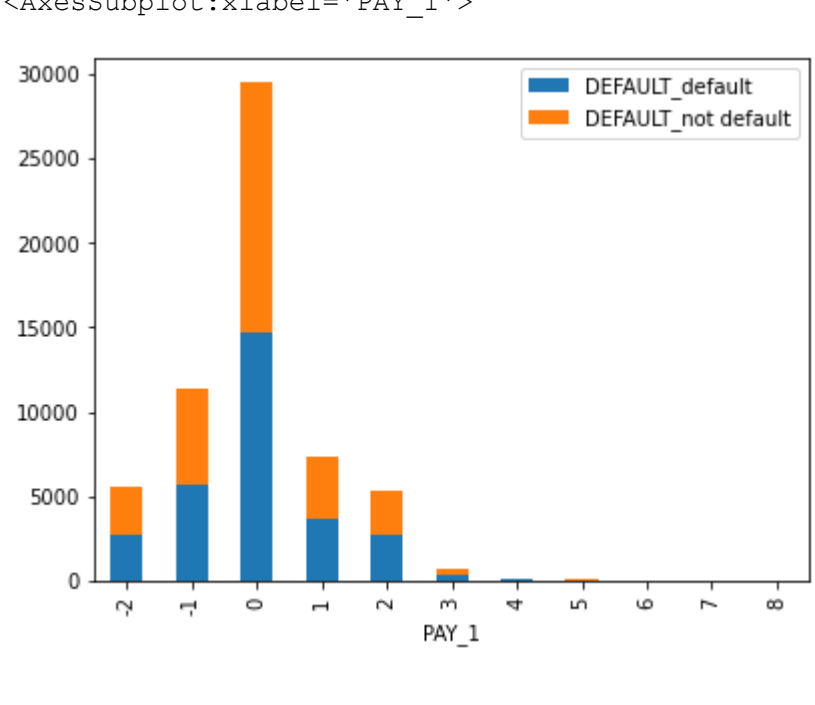
Looking into PAY_1 vs. default

In [37]:
```
plt.hist(df['PAY_1'])
plt.show()
```



In [38]:
```
df[['PAY_1','DEFAULT_default','DEFAULT_not default']].groupby('PAY_1').count().plot.bar(stacked=True)
#This isn't a useful plot.  Just telling counts of PAY1 groups
```

Out[38]:
```
<AxesSubplot:xlabel='PAY_1'>
```



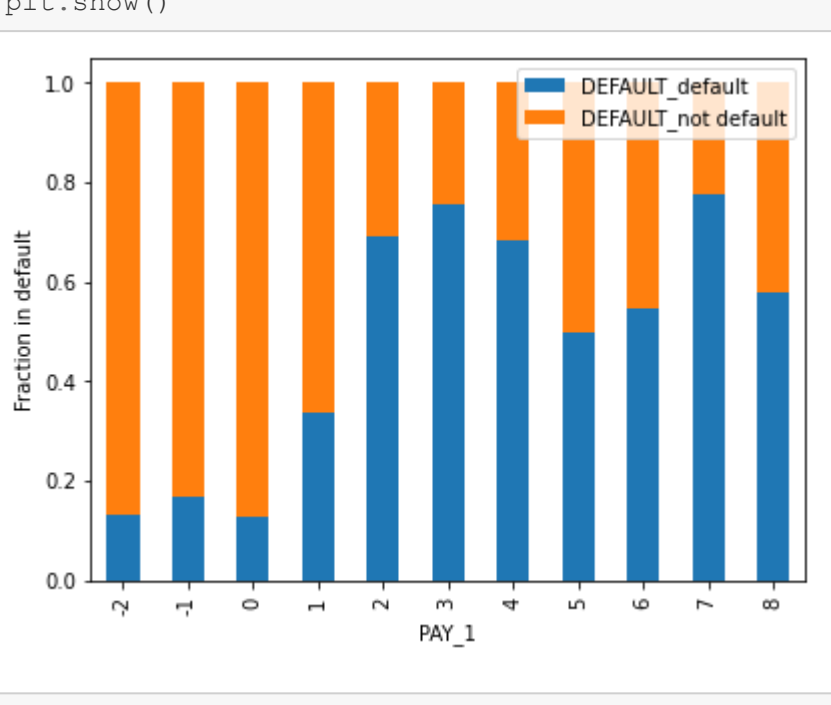
Consider doing an sum of all the PAY ones in a similar fashion

In [39]:
```
df[['PAY_1','DEFAULT_default','DEFAULT_not default']].groupby('PAY_1').mean()
```

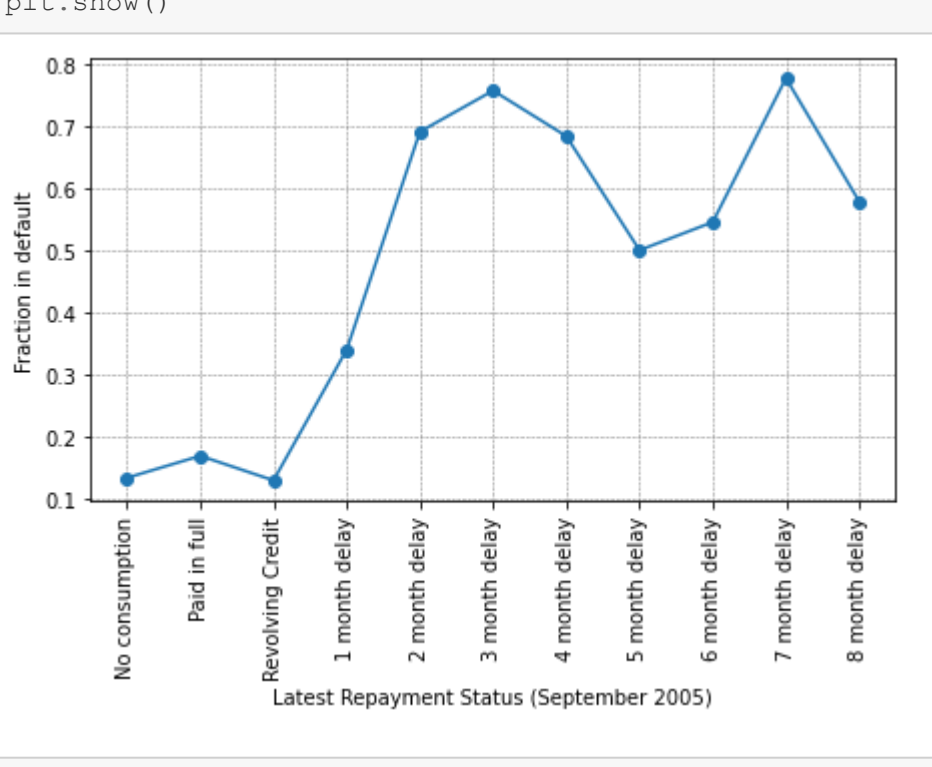
Out[39]:

PAY_1	DEFAULT_default	DEFAULT_not default
-2	0.132294	0.867706
-1	0.167781	0.832219
0	0.128113	0.871887
1	0.339479	0.660521
2	0.691414	0.308586
3	0.757764	0.242236
4	0.684211	0.315789
5	0.500000	0.500000
6	0.545455	0.454545
7	0.777778	0.222222
8	0.578947	0.421053

In [40]:
```
df[['PAY_1','DEFAULT_default','DEFAULT_not default']].groupby('PAY_1').mean().plot(kind='bar',stacked=True, rot=90)
plt.ylabel('Fraction in default')
#plt.xticks(range(-2,9,1), PAY_XLABELS)
#plt.xticks(range(0,11,1), PAY_XLABELS) #for some reason, adding xticks as bar forces the axis to start at "0"
#plt.legend((0,1), ('label1', 'label2', 'label3'))
#plt.savefig('PAY_1_bar.png')
plt.show()
```



In [41]:
```
df[['PAY_1','DEFAULT_default']].groupby('PAY_1').mean().plot(kind='line', rot=90, style='o-',legend=None)
plt.ylabel('Fraction in default')
plt.xticks(range(-2,9,1), PAY_XLABELS)
plt.grid(color='gray', linestyle='--', linewidth=0.5)
plt.xlabel('Latest Repayment Status (September 2005)')
plt.tight_layout()
plt.savefig('PAY_1.png')
plt.show()
```



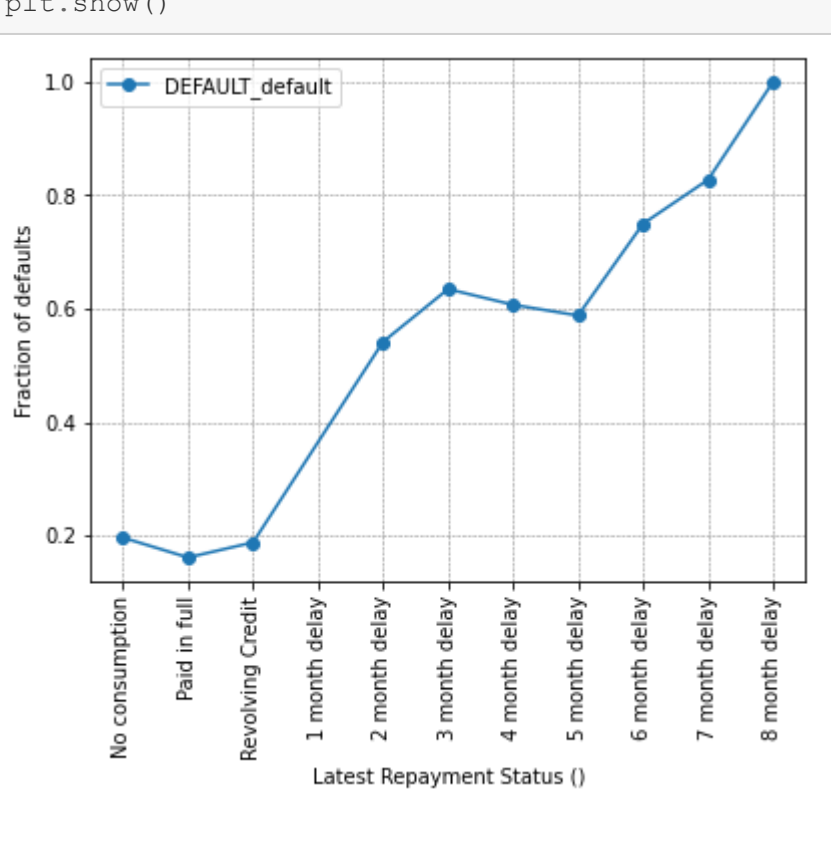
In [42]:
```
print("First number is and second number is {}".format(df['DEFAULT_default'].corr(df['PAY_1'])))
```

```
First number is and second number is 0.32479372847862265
```

In [43]:
```
print("PAY_1: {}".format(df['DEFAULT_default'].corr(df['PAY_1'])))
print("PAY_2: {}".format(df['DEFAULT_default'].corr(df['PAY_2'])))
print("PAY_3: {}".format(df['DEFAULT_default'].corr(df['PAY_3'])))
print("PAY_4: {}".format(df['DEFAULT_default'].corr(df['PAY_4'])))
print("PAY_5: {}".format(df['DEFAULT_default'].corr(df['PAY_5'])))
```

```
PAY_1: 0.32479372847862265
PAY_2: 0.26355120167216467
PAY_3: 0.235252513724916
PAY_4: 0.2166136368424245
PAY_5: 0.2041489138761666
```

In [44]:
```
df[['PAY_5','DEFAULT_default']].groupby('PAY_5').mean().plot(kind='line', rot=90, style='o-')
plt.ylabel('Fraction of defaults')
plt.xticks(range(-2,9,1), PAY_XLABELS)
plt.grid(color='gray', linestyle='--', linewidth=0.5)
plt.xlabel('Latest Repayment Status ()')
plt.show()
```



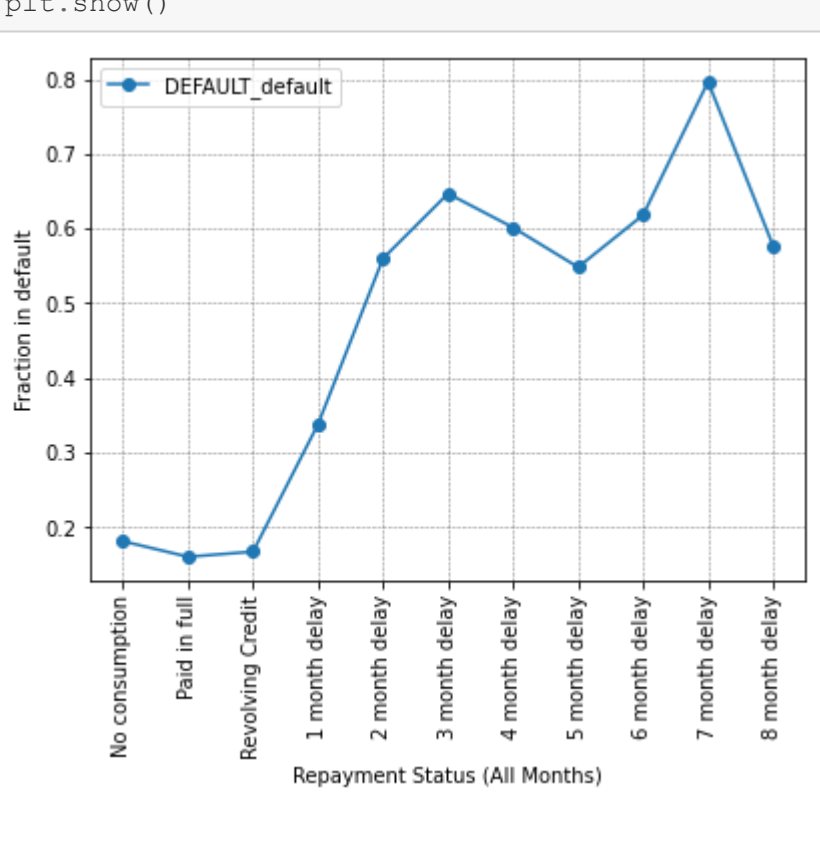
Perhaps try stacking all of the PAY_X columns for a combined result?

In [45]:
```
df1=df.melt(id_vars=['DEFAULT_default'], value_vars=['PAY_1','PAY_2','PAY_3','PAY_4','PAY_5'], value_name='PAY_ALL')
df1.head()
```

Out[45]:

	DEFAULT_default	variable	PAY_ALL
0	1	PAY_1	2
1	1	PAY_1	-1
2	0	PAY_1	0
3	0	PAY_1	0
4	0	PAY_1	-1

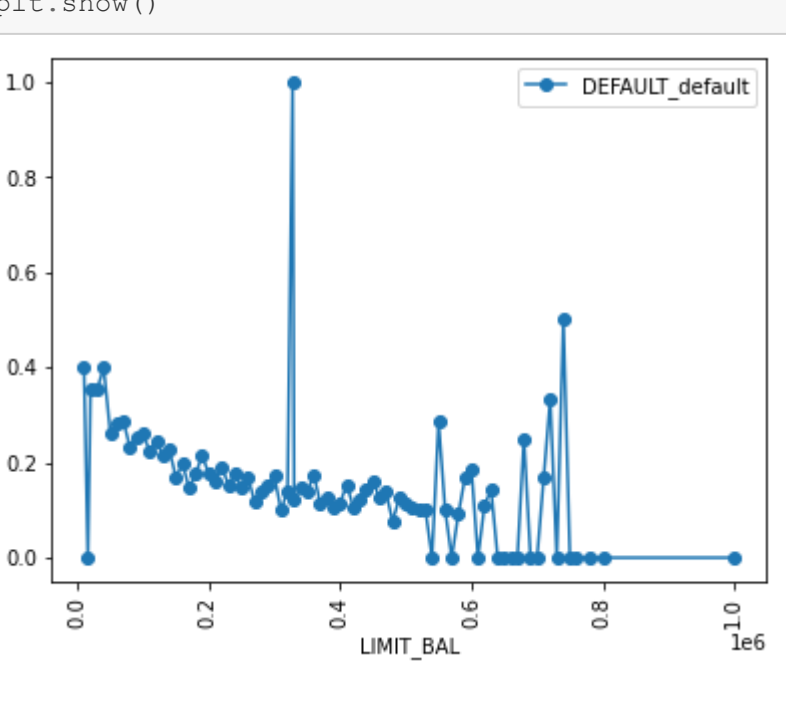
In [46]:
```
df1[['PAY_ALL','DEFAULT_default']].groupby('PAY_ALL').mean().plot(kind='line', rot=90, style='o-')
plt.ylabel('Fraction in default')
plt.xticks(range(-2,9,1), PAY_XLABELS)
plt.grid(color='gray', linestyle='--', linewidth=0.5)
plt.xlabel('Repayment Status (All Months)')
plt.show()
```



Not more insightful by ganging them all together. It was worth a shot

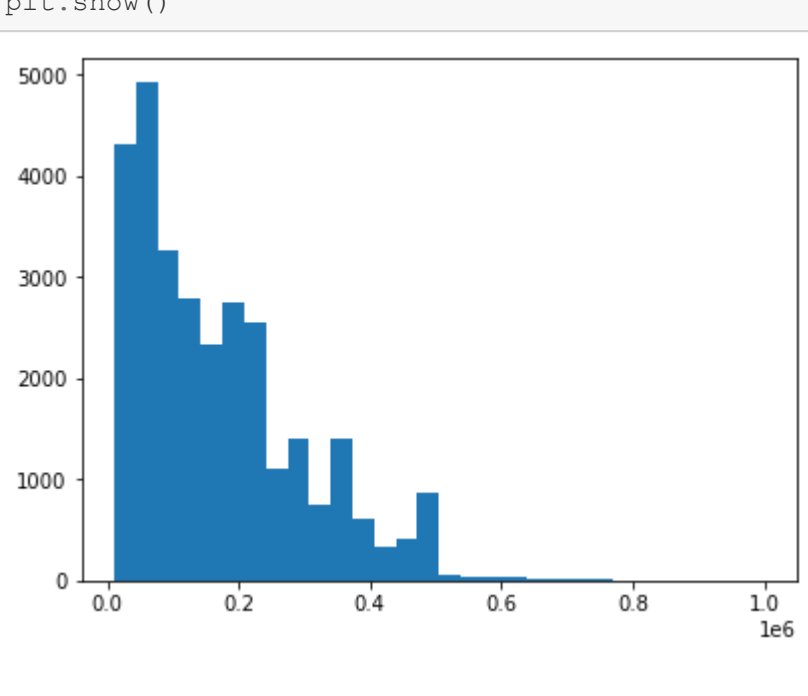
Looking into how limit balance impacts rate of default

In [47]:
```
df[['LIMIT_BAL','DEFAULT_default']].groupby('LIMIT_BAL').mean().plot(kind='line', rot=90, style='o-')
plt.show()
```



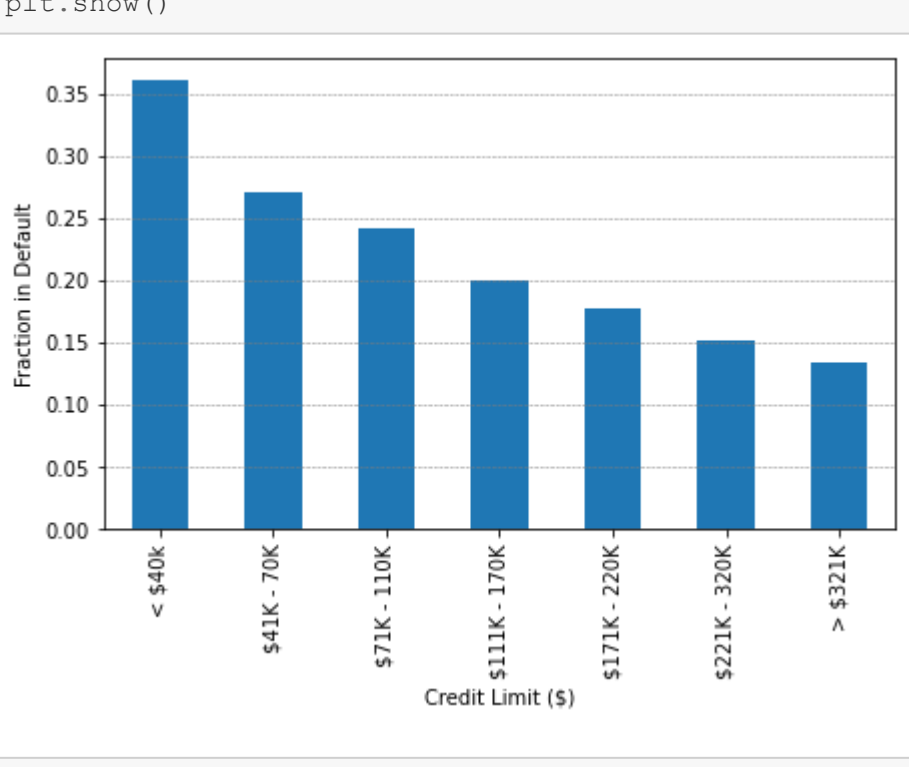
I bet it gets noisy at the end because of fewer customers with LIMIT_BAL that high. Let's check

In [48]:
```
plt.hist(df['LIMIT_BAL'], bins=30)
plt.show()
```



I'll try to make the prior one as a bar to blend a bit. Will need to discretize

In [49]:
```
#make a subset df with the columns we care about
df2 = df.loc[:,['LIMIT_BAL','DEFAULT_default']]
#discretize
bins=7
df2.loc[:,'LimitBin']=pd.qcut(df2.LIMIT_BAL,bins)
#get the mean default value (averaged rate) by bin
df2.groupby('LimitBin').mean().plot.bar(y='DEFAULT_default', legend=None)
plt.grid(axis='y',color='gray', linestyle='--', linewidth=0.5)
plt.xticks(np.arange(7), ('< $40k', '$41K - 70K', '$71K - 110K', '$111K - 170K', '$171K - 220K', '$221K - 320K', '> $321K'))
plt.xlabel('Credit Limit ($)')
plt.ylabel('Fraction in Default')
plt.tight_layout()
plt.savefig('credit_limit.png')
plt.show()
```



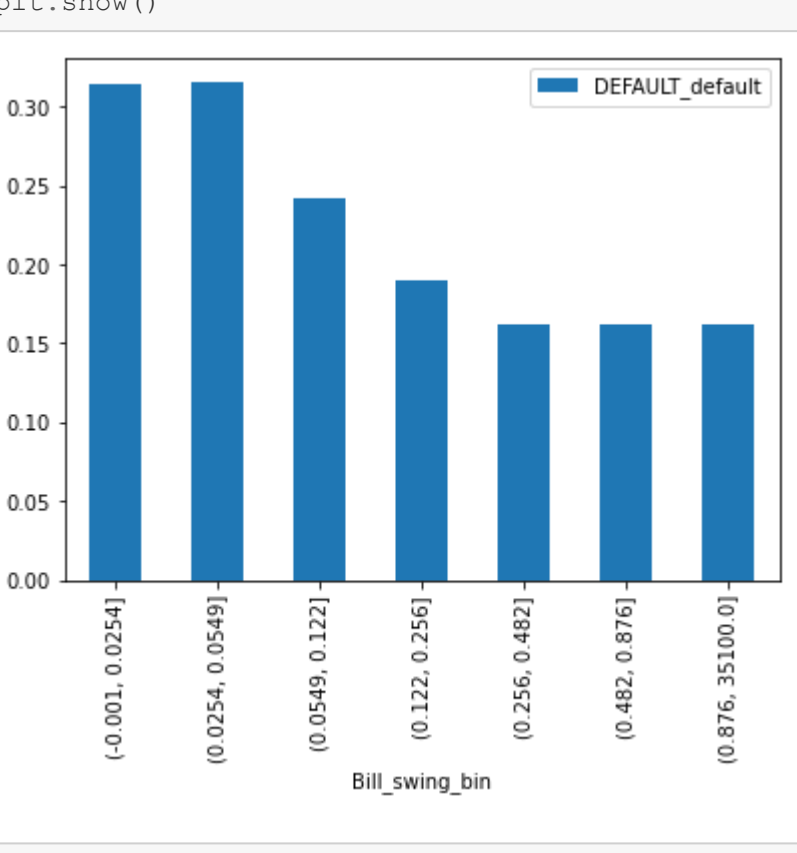
In [50]:
```
df2['DEFAULT_default'].corr(df2['LIMIT_BAL'])
```

Out[50]:
```
-0.15351987639350764
```

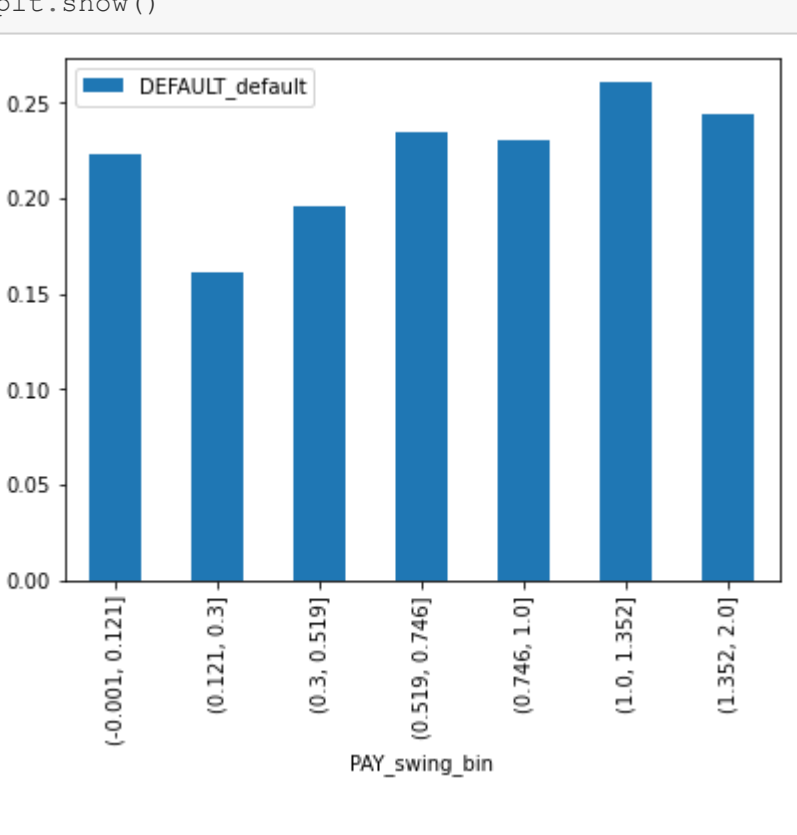
Looking into swings in bills and payments

cumulative add of % change every month

In [51]:
```
df3 = df.copy()
#Fails really hard for monthly bills of $0 or really low when doing as a % change.
  # instead, do as a weighted delta.  Could do as a for loop with indexes, but would be hard to read later
df3.loc[:,'Bill_swing'] = (abs(df3.BILL_AMT1 - df3.BILL_AMT2) + \
                          abs(df3.BILL_AMT2 - df3.BILL_AMT3) + \
                          abs(df3.BILL_AMT3 - df3.BILL_AMT4) + \
                          abs(df3.BILL_AMT4 - df3.BILL_AMT5) + \
                          abs(df3.BILL_AMT5 - df3.BILL_AMT6)) / (abs(df3.iloc[:, 10:16].sum(axis=1)+.1))
#had to add a +.1 to avoid divide by 0 errors
#as before, discretize, group bins, look at mean Defaulting rate
bins=7
df3.loc[:,'Bill_swing_bin']=pd.qcut(df3.Bill_swing,bins)
df3.groupby('Bill_swing_bin').mean().plot.bar(y='DEFAULT_default')
plt.show()
```



In [52]:
```
df3.loc[:,'PAY_swing'] = (abs(df3.PAY_AMT1 - df3.PAY_AMT2) + \
                         abs(df3.PAY_AMT2 - df3.PAY_AMT3) + \
                         abs(df3.PAY_AMT3 - df3.PAY_AMT4) + \
                         abs(df3.PAY_AMT4 - df3.PAY_AMT5) + \
                         abs(df3.PAY_AMT5 - df3.PAY_AMT6)) / (abs(df3.iloc[:, 16:22].sum(axis=1)+.1))
#had to add a +.1 to avoid divide by 0 errors
#as before, discretize, group bins, look at mean Defaulting rate
bins=7
df3.loc[:,'PAY_swing_bin']=pd.qcut(df3.PAY_swing,bins)
df3.groupby('PAY_swing_bin').mean().plot.bar(y='DEFAULT_default')
plt.show()
```



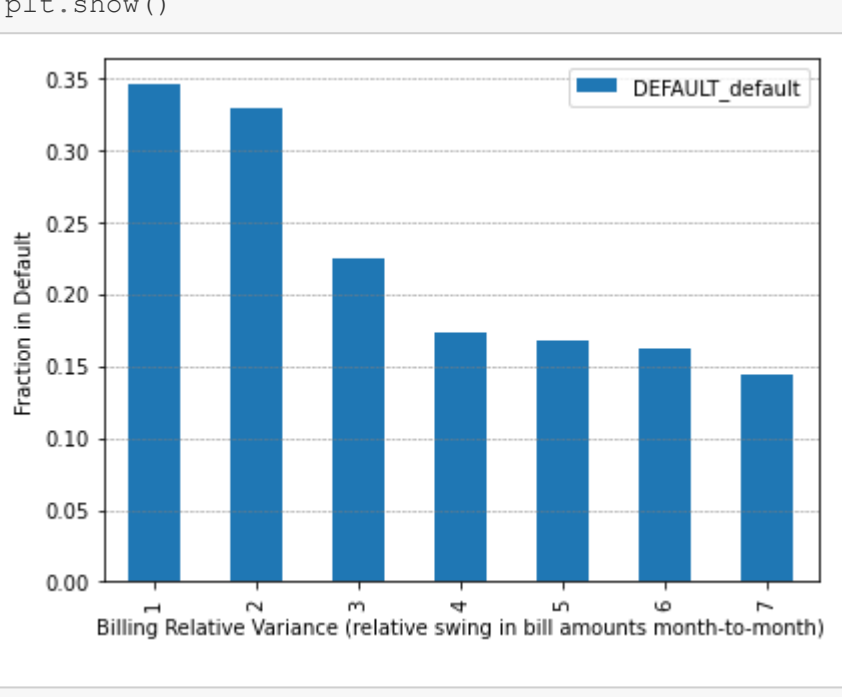
Should probably just do this as relative variance over the columns

It's a more known analysis method, and pretty similar

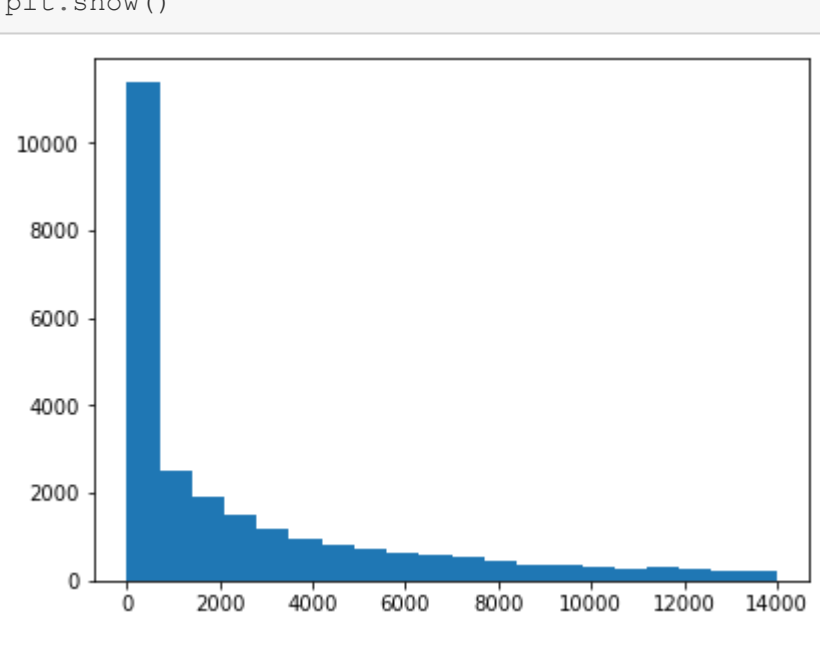
In [53]:
```
df3['Bill_swing']
```

Out[53]:
```
0        0.507911
1        0.168705
2        0.173826
3        0.105955
4        0.455610
           ...
29995    0.292000
29996    0.768337
29997    0.290059
29998    0.688782
29999    0.157168
Name: Bill_swing, Length: 30000, dtype: float64
```

In [55]:
```
#variance of columns divided by mean
# +0.1 to avoid div/0
df3.loc[:,'Bill_relvar']=df3.iloc[:, 10:16].var(axis=1)/ (df3.iloc[:, 10:16].mean(axis=1)+.1)
#as before, discretize, group bins, look at mean Defaulting rate
bins=7
BinLabels=[1,2,3,4,5,6,7]
#df3.loc[:,'Bill_relvar_bin']=pd.qcut(df3.Bill_relvar,bins)
df3.loc[:,'Bill_relvar_bin']=pd.qcut(df3.Bill_relvar,bins,labels=BinLabels)
df3.groupby('Bill_relvar_bin').mean().plot.bar(y='DEFAULT_default')
plt.grid(axis='y',color='gray', linestyle='--', linewidth=0.5)
#plt.xticks(np.arange(7), ('< $40k', '$41K - 70K', '$71K - 110K', '$111K - 170K', '$171K - 220K', '$221K - 320K', '> $321K'))
plt.xlabel('Billing Relative Variance (relative swing in bill amounts month-to-month)')
plt.ylabel('Fraction in Default')
plt.show()
```



In [56]:
```
plt.hist(df3['Bill_relvar'], bins=20, range=(0, 14000))
plt.show()
```



This is really surprising. The ones with wild swings in their bill amount tend to default less

In [57]:
```
df3['DEFAULT_default'].corr(df3['Bill_relvar'])
```

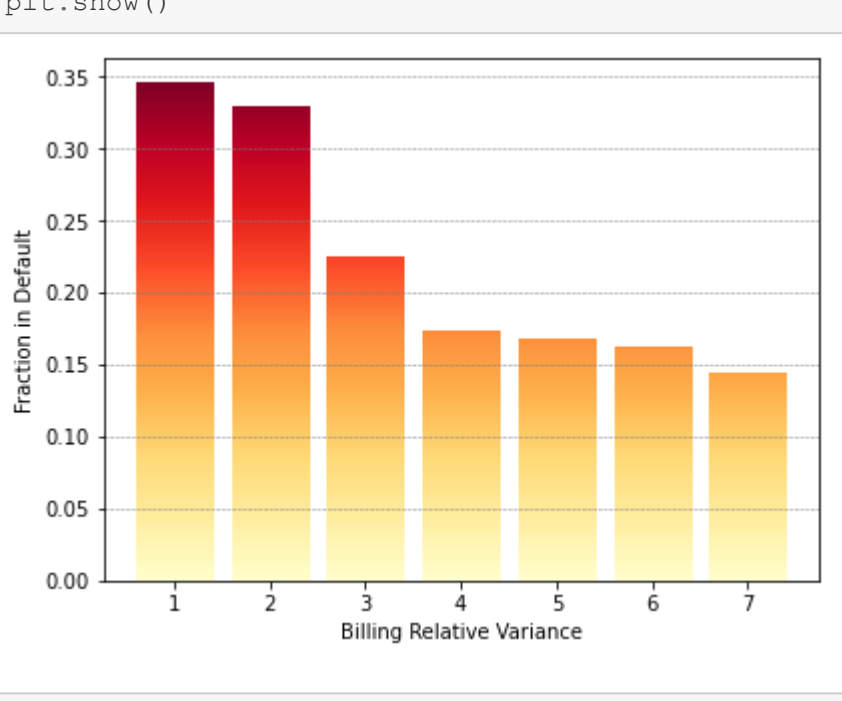
Out[57]:
```
-0.026200868279733114
```

That's a surprisingly low correlation given the clear dependency on the plot.

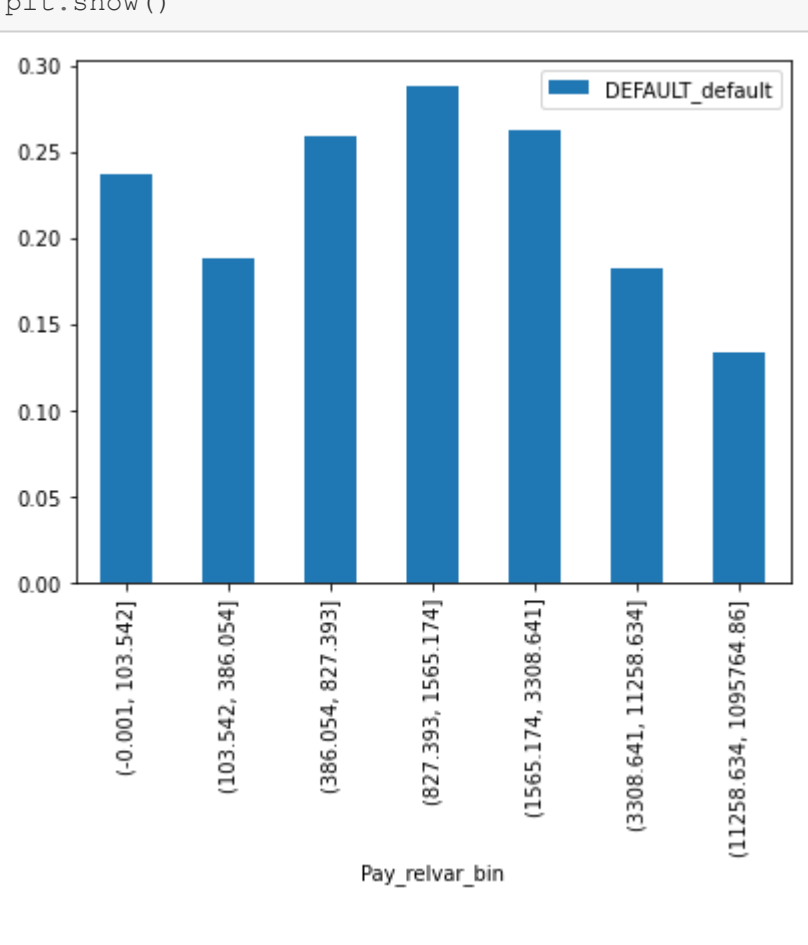
This is a key plot - let's make it pretty

In [58]:
```
#fig, ax = plt.subplots()
#bar = ax.bar([1,2,3,4,5,6],[4,5,6,3,7,5])

def gradientbars(bars,ydata):
    #grad = np.atleast_2d(np.linspace(0,1,256)).T
    ax = bars[0].axes
    lim = ax.get_xlim()+ax.get_ylim()
    for bar in bars:
        bar.set_zorder(1)
        bar.set_facecolor("none")
        x,y = bar.get_xy()
        w, h = bar.get_width(), bar.get_height()
        grad = np.atleast_2d(np.linspace(0,1*h/max(ydata),256)).T
        ax.imshow(grad, extent=[x,x+w,y,y+h], origin='lower',aspect="auto",zorder=0, norm=cm.colors.NoNorm(vmin=0,vmax=1),cmap=plt.get_cmap('YlOrRd'))
    ax.axis(lim)

#gradientbars(bar, [4,5,6,3,7,5])
```

In [59]:
```
B_relvar_df = df3.groupby('Bill_relvar_bin', as_index=False).mean()
xplt = B_relvar_df.Bill_relvar_bin.array
yplt = B_relvar_df.DEFAULT_default.array
fig, ax = plt.subplots()
bar3 = ax.bar(xplt, yplt)
plt.grid(axis='y',color='gray', linestyle='--', linewidth=0.5)
plt.xlabel('Billing Relative Variance')
plt.ylabel('Fraction in Default')
gradientbars(bar3,yplt)
plt.savefig('bill_relvar.png')
plt.show()
```



In [60]:
```
#variance of columns divided by mean
# +0.1 to avoid div/0
df3.loc[:,'Pay_relvar']=df3.iloc[:, 16:22].var(axis=1)/ (df3.iloc[:, 16:22].mean(axis=1)+.1)
#as before, discretize, group bins, look at mean Defaulting rate
bins=7
df3.loc[:,'Pay_relvar_bin']=pd.qcut(df3.Pay_relvar,bins)
df3.groupby('Pay_relvar_bin').mean().plot.bar(y='DEFAULT_default')
plt.show()
```



Let's look for anything else... gender perhaps

In [61]:
```
df.loc[df.SEX_female==1].DEFAULT_default.mean()
```

Out[61]:
```
0.20776280918727916
```

In [62]:
```
df.loc[df.SEX_male==1].DEFAULT_default.mean()
```

Out[62]:
```
0.2416722745625841
```

Men are 4% more likely to default

In [63]:
```
df.loc[df.SEX_female==1].LIMIT_BAL.mean()
```

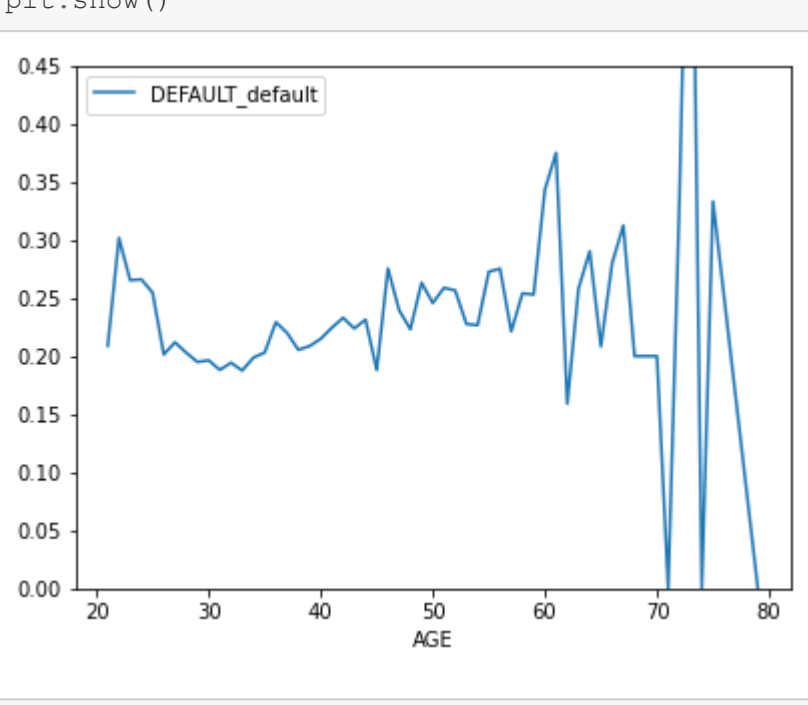
Out[63]:
```
170086.46201413427
```

In [64]:
```
df.loc[df.SEX_male==1].LIMIT_BAL.mean()
```

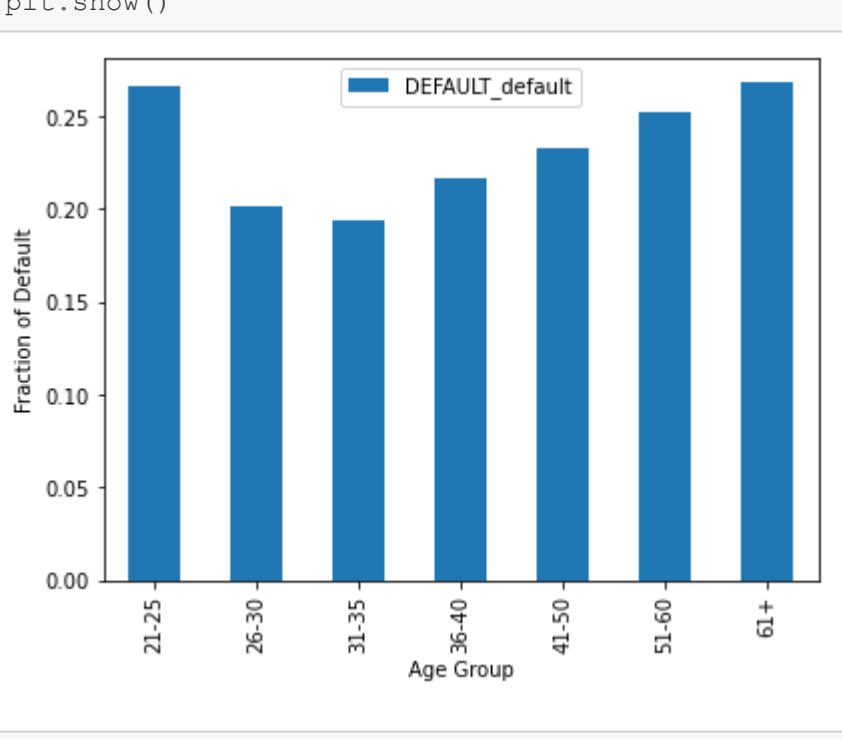
Out[64]:
```
163519.8250336474
```

Pretty matched on credit limit vs. gender

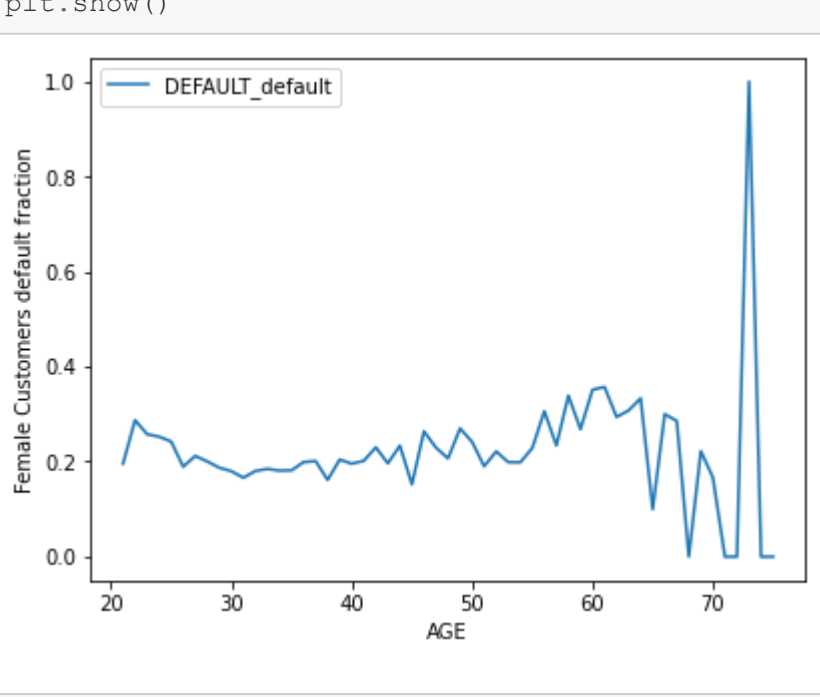
In [65]:
```
df.groupby('AGE').mean().plot.line(y='DEFAULT_default')
plt.ylim(0,0.45)
plt.show()
```



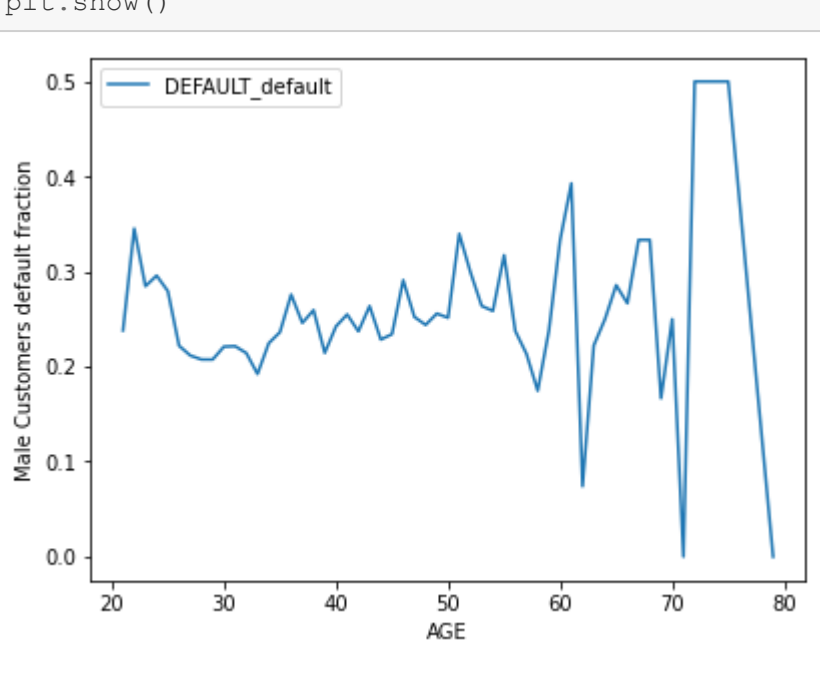
In [66]:
```
df_age=df.loc[:,['AGE','SEX_female','SEX_male','DEFAULT_default','DEFAULT_not default']]
ageBinCuts = [20,25,30,35,40,50,60,100]
plt_ageBinLabels = ['21-25','26-30','31-35','36-40','41-50','51-60','61+']
ageBinLabels = [1,2,3,4,5,6,7]
ageBin = pd.cut(df_age['AGE'],bins=ageBinCuts, labels=ageBinLabels)
df_age.loc[:,'ageBin']=ageBin
df_age.groupby('ageBin').mean().plot.bar(y='DEFAULT_default')
plt.xticks([0,1,2,3,4,5,6],plt_ageBinLabels)
plt.xlabel('Age Group')
plt.ylabel('Fraction of Default')
plt.show()
```



In [67]:
```
df_female = df.loc[df.SEX_female==1]
df_female.groupby('AGE').mean().plot.line(y='DEFAULT_default')
plt.ylabel('Female Customers default fraction')
plt.show()
```

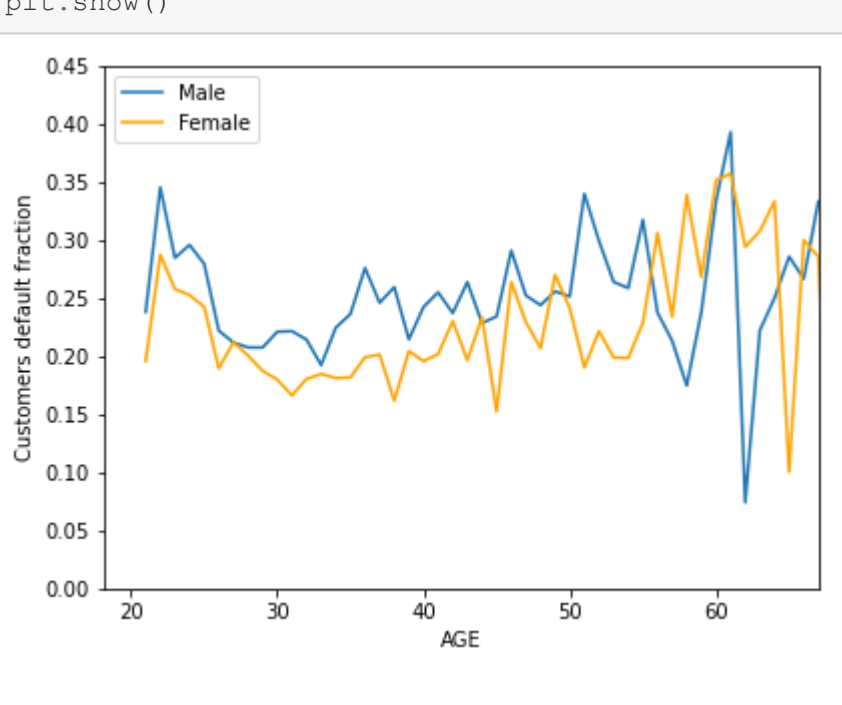


In [68]:
```
df_male = df.loc[df.SEX_male==1]
df_male.groupby('AGE').mean().plot.line(y='DEFAULT_default')
plt.ylabel('Male Customers default fraction')
plt.show()
```



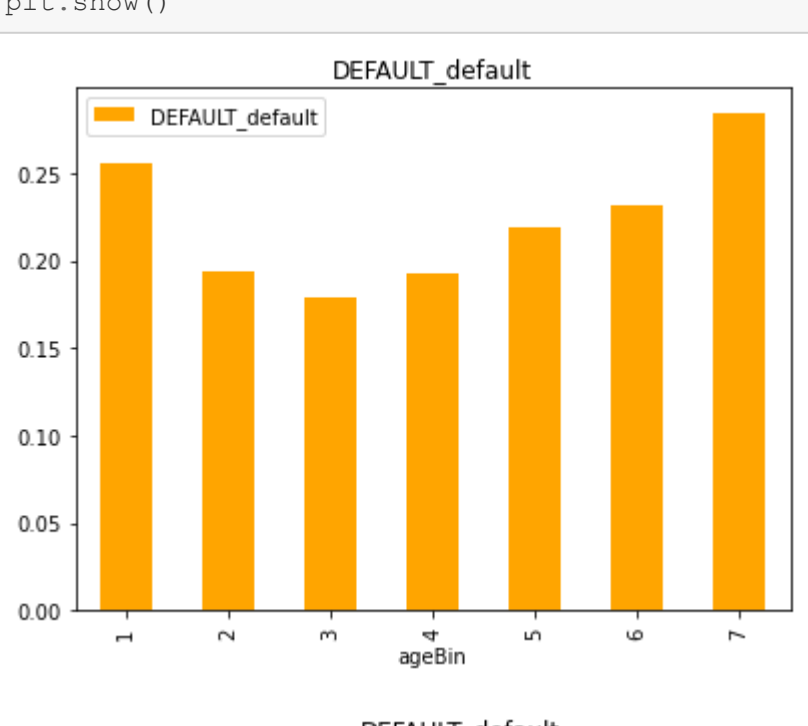
Let's overlay these for male and female, and plot in seaborn

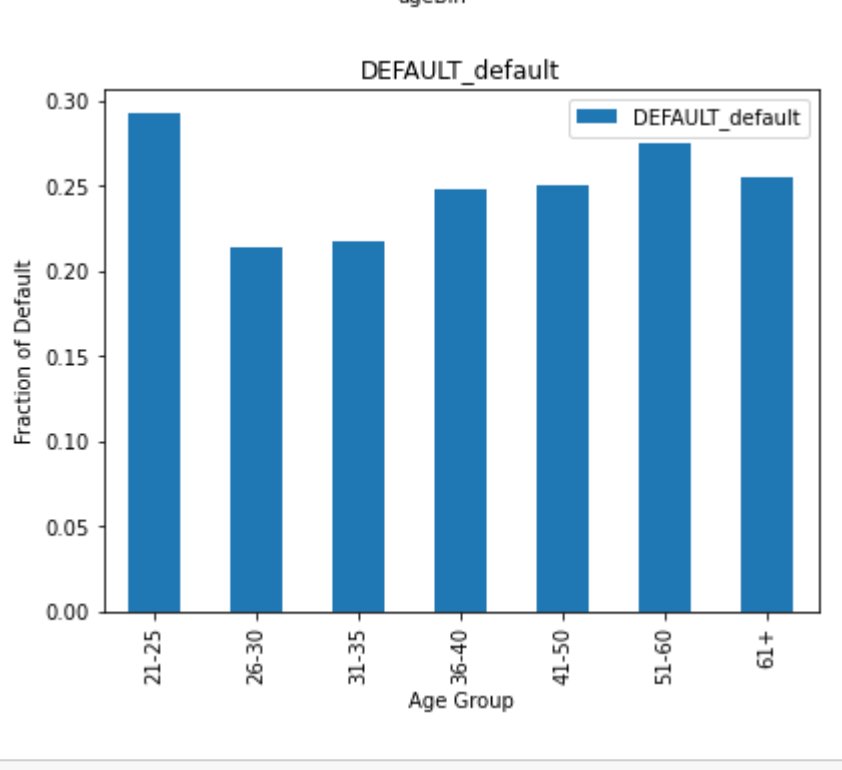
In [69]:
```
ax = plt.gca()
df_male = df.loc[df.SEX_male==1]
df_male.groupby('AGE').mean().plot.line(y='DEFAULT_default', ax=ax)
df_female = df.loc[df.SEX_female==1]
df_female.groupby('AGE').mean().plot.line(y='DEFAULT_default', ax=ax, color='orange')
plt.ylabel('Customers default fraction')
plt.legend(['Male', 'Female'])
plt.xlim(right=67)
plt.ylim(0,0.45)
plt.show()
```



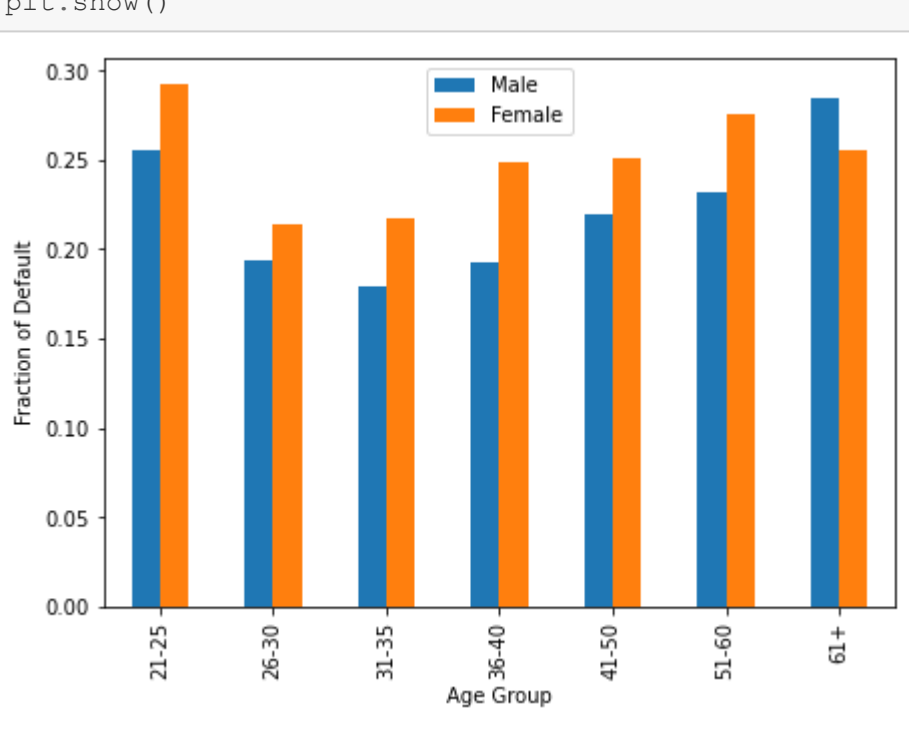
There seems to be a bit of a dip around age=30 for both genders. This may look better as a binned bar chart

In [70]:
```
df_age_f=df_age.loc[df_age.SEX_female==1]
df_age_f.groupby('ageBin').mean().plot.bar(y='DEFAULT_default', subplots=True, color='orange')
df_age_m=df_age.loc[df_age.SEX_male==1]
df_age_m.groupby('ageBin').mean().plot.bar(y='DEFAULT_default', subplots=True)
plt.xticks([0,1,2,3,4,5,6],plt_ageBinLabels)
plt.xlabel('Age Group')
plt.ylabel('Fraction of Default')
plt.show()
```





In [71]:
```
df_temp1=df_age_f.groupby('ageBin').mean()
df_temp2=df_age_m.groupby('ageBin').mean()
#combine the default columns
df_temp1.rename(columns={'DEFAULT_default':'Default_Female'}, inplace=True)
df_temp1.loc[:,'Default_Male']=df_temp2.DEFAULT_default
df_temp1.plot.bar(y=['Default_Female','Default_Male'])
plt.xticks([0,1,2,3,4,5,6],plt_ageBinLabels)
plt.xlabel('Age Group')
plt.ylabel('Fraction of Default')
plt.legend(['Male','Female'])
plt.tight_layout()
plt.savefig('age_gender_bar.png')
plt.show()
```



In [72]:
```
df_male.groupby('AGE').mean().shape
```

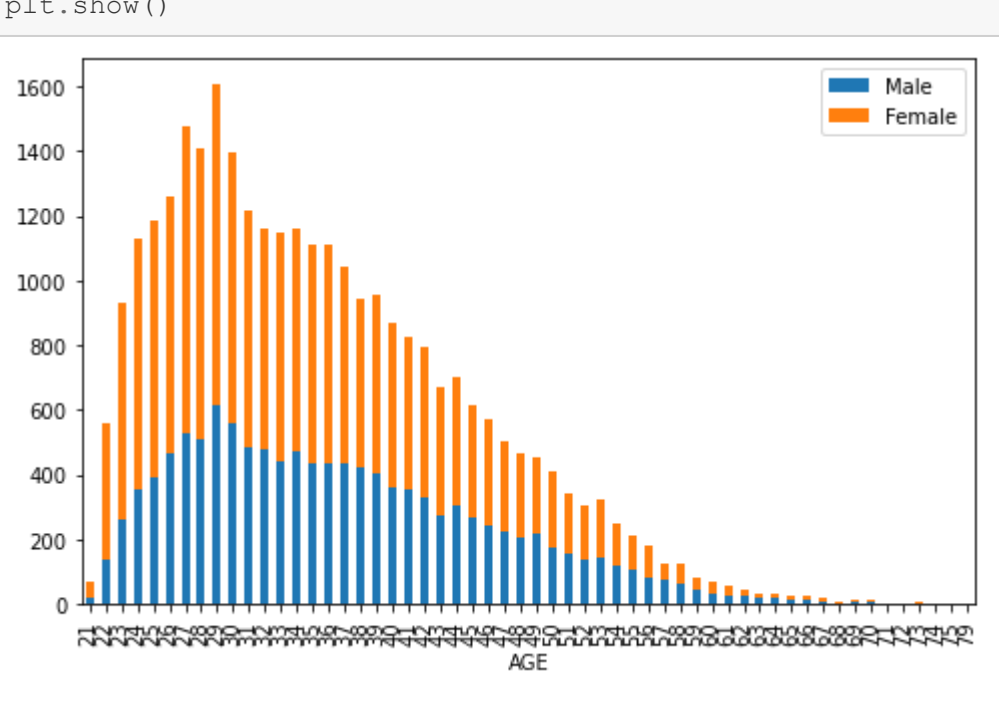
Out[72]:
```
(55, 29)
```

In [73]:
```
df4 = df[['AGE','SEX_female','SEX_male']]
```

In [74]:
```
df4.groupby('AGE')['SEX_female'].value_counts()
```

Out[74]:
```
AGE  SEX_female
21   1              46
     0              21
22   1             421
     0             139
23   1             671
                  ...
73   1               2
74   1               1
75   0               2
     1               1
79   0               1
Name: SEX_female, Length: 110, dtype: int64
```

In [75]:
```
df4.groupby('AGE')['SEX_female'].value_counts().unstack(level=1).plot.bar(stacked=True, figsize=[8,5])
plt.legend(['Male','Female'])
plt.show()
```



Let's take a quick look at education to see if there's anything there. Heatmap says no....

In [76]:
```
hs = df.loc[df['EDUCATION_high school']==1].DEFAULT_default.mean()
uni = df.loc[df['EDUCATION_university']==1].DEFAULT_default.mean()
gs = df.loc[df['EDUCATION_graduate school']==1].DEFAULT_default.mean()
oth = df.loc[df['EDUCATION_other']==1].DEFAULT_default.mean()
print("high school: {}".format(hs))
print("university: {}".format(uni))
print("graduate school: {}".format(gs))
print("other: {}".format(oth))
```

```
high school: 0.2515761643278422
university: 0.23734853884533144
graduate school: 0.19234766178554558
other: 0.07051282051282051
```

In [77]:
```
education_df = pd.DataFrame({'Education':['high school', 'university','graduate school','other'],
                             'Default':[hs,uni,gs,oth]})
education_df.head()
```

Out[77]:

	Education	Default
0	high school	0.251576
1	university	0.237349
2	graduate school	0.192348
3	other	0.070513

In [78]:
```
df.loc[df['EDUCATION_other']==1].shape
```

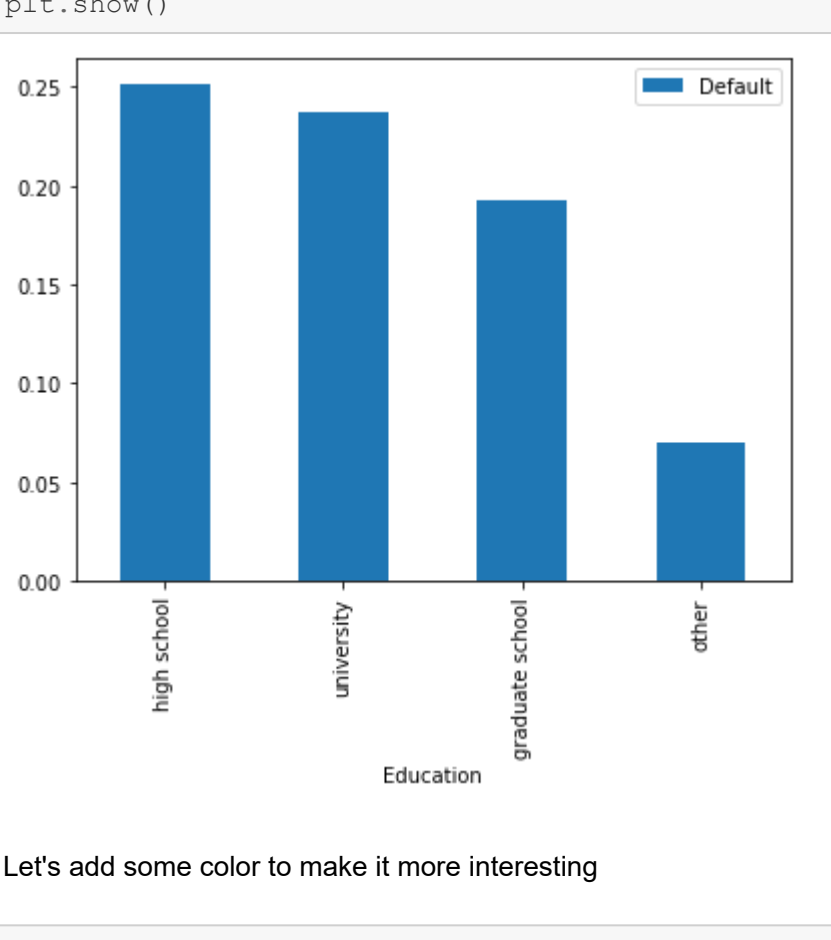
Out[78]:
```
(468, 30)
```

In [79]:
```
df.shape
```

Out[79]:
```
(30000, 30)
```

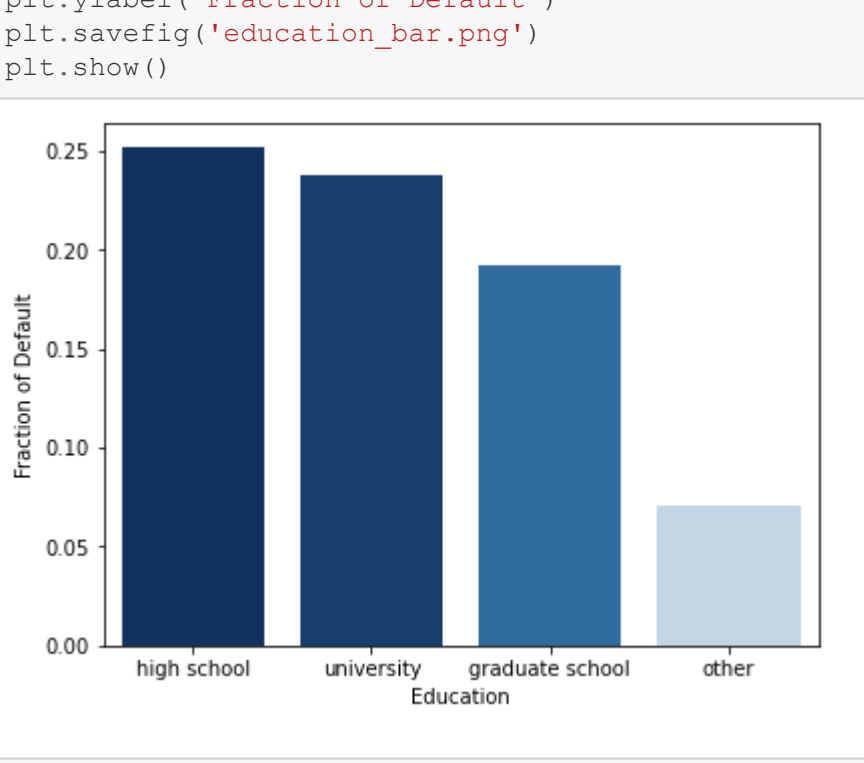
Um... that seems pretty significant for 'Other' - though only 468 of them out of 30000! 1.5% of population

In [80]:
```
education_df.plot.bar(x='Education', y='Default')
plt.show()
```



Let's add some color to make it more interesting

In [81]:
```
sns.barplot(x=education_df['Education'], y=education_df['Default'], palette=cm.Blues(education_df['Default']*4))
plt.ylabel('Fraction of Default')
plt.savefig('education_bar.png')
plt.show()
```



In [ ]: