In [1]:

```
import pandas as pd
import matplotlib.pyplot as ply
import seaborn as sns
import numpy as np
import scipy.stats as st
import statsmodels.formula.api as smf
```

In [2]:

```
import pandas as pd
import matplotlib.pyplot as ply
import seaborn as sns
import numpy as np
import scipy.stats as st
# Load libraries
import pandas as pd
#import numpy
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn import linear_model
import statsmodels.formula.api as smf
from sklearn import linear_model
from patsy.builtins import Q
from sklearn.linear_model import RidgeCV
```

In [3]:

```
df = pd.read_csv("LoanStats3b.csv",low_memory =False)
```

In [4]:

```
#df.head(10)
```

In [5]:

```
print('Number of missing values per column:')
countMissing = df.isnull().sum()
print(countMissing)
```

```
Number of missing values per column:
```

```
id                                      188181
member_id                               188181
loan_amnt                                    0
funded_amnt                                  0
funded_amnt_inv                              0
term                                         0
int_rate                                     0
installment                                  0
grade                                        0
sub_grade                                    0
emp_title                                11737
emp_length                                7887
home_ownership                               0
annual_inc                                   0
verification_status                          0
issue_d                                      0
loan_status                                  0
pymnt_plan                                   0
url                                     188181
desc                                    106703
purpose                                      0
title                                        7
zip_code                                     0
addr_state                                   0
dti                                          0
delinq_2yrs                                  0
earliest_cr_line                             0
inq_last_6mths                               0
mths_since_last_delinq                  107573
mths_since_last_record                  170707
                                           ...
sec_app_mort_acc                        188181
sec_app_open_acc                        188181
sec_app_revol_util                      188181
sec_app_open_act_il                     188181
sec_app_num_rev_accts                   188181
sec_app_chargeoff_within_12_mths        188181
sec_app_collections_12_mths_ex_med      188181
sec_app_mths_since_last_major_derog     188181
hardship_flag                                0
hardship_type                           188101
hardship_reason                         188101
hardship_status                         188101
deferral_term                           188101
hardship_amount                         188101
```

```
hardship_start_date                                188101
hardship_end_date                                  188101
payment_plan_start_date                            188101
hardship_length                                    188101
hardship_dpd                                       188101
hardship_loan_status                               188101
orig_projected_additional_accrued_interest         188112
hardship_payoff_balance_amount                     188101
hardship_last_payment_amount                       188101
debt_settlement_flag                                    0
debt_settlement_flag_date                          186103
settlement_status                                  186103
settlement_date                                    186103
settlement_amount                                  186103
settlement_percentage                              186103
settlement_term                                    186103
Length: 144, dtype: int64
```

In [6]:

```
# Drop the columns where all elements are missing
#df = df.dropna(axis=0,how='any')
df = df.dropna(axis=1,how='all')
```

In [7]:

```
# Columns that are being dropped
df=df.drop(['hardship_dpd','hardship_loan_status','emp_title','h
ardship_type','hardship_reason','hardship_status','deferral_term
','hardship_amount'],axis=1)
df=df.drop(['settlement_status','settlement_date','settlement_am
ount','settlement_term', 'policy_code', 'acc_now_delinq', 'num_t
l_30dpd'],axis=1)
df=df.drop(['hardship_start_date','hardship_end_date','orig_proj
ected_additional_accrued_interest','hardship_payoff_balance_amou
nt','debt_settlement_flag_date'],axis=1)
df=df.drop(['total_bal_ex_mort', 'num_sats', 'tot_cur_bal', 'tax
_liens', 'zip_code','addr_state','title', 'num_tl_90g_dpd_24m'],
axis=1)
df=df.drop(['payment_plan_start_date','hardship_length','hardshi
p_last_payment_amount','settlement_percentage', 'collections_12_
mths_ex_med'],axis=1)
df=df.drop(['mths_since_last_record', 'mths_since_recent_bc_dlq'
, 'mths_since_recent_revol_delinq', 'mths_since_last_delinq' ,'m
ths_since_last_major_derog'],axis=1)
df=df.drop(['hardship_flag', 'debt_settlement_flag', 'num_tl_120
dpd_2m', 'chargeoff_within_12_mths', 'delinq_amnt', 'application
_type', 'emp_length'],axis=1)
df=df.drop(['out_prncp', 'out_prncp_inv'], axis=1)
```

In [8]:

```
print('Number of missing values per column:')
countMissing = df.isnull().sum()
print(countMissing)
```

```
Number of missing values per column:
loan_amnt                       0
funded_amnt                     0
funded_amnt_inv                 0
term                            0
int_rate                        0
installment                     0
grade                           0
sub_grade                       0
home_ownership                  0
annual_inc                      0
verification_status             0
```

```
issue_d                       0
loan_status                   0
pymnt_plan                    0
desc                     106703
purpose                       0
dti                           0
delinq_2yrs                   0
earliest_cr_line              0
inq_last_6mths                0
open_acc                      0
pub_rec                       0
revol_bal                     0
revol_util                  125
total_acc                     0
initial_list_status           0
total_pymnt                   0
total_pymnt_inv               0
total_rec_prncp               0
total_rec_int                 0
                            ...
last_credit_pull_d           10
tot_coll_amt              27741
total_rev_hi_lim          27741
acc_open_past_24mths       7495
avg_cur_bal               27747
bc_open_to_buy             9025
bc_util                    9112
mo_sin_old_il_acct        33872
mo_sin_old_rev_tl_op      27742
mo_sin_rcnt_rev_tl_op     27742
mo_sin_rcnt_tl            27741
mort_acc                   7495
mths_since_recent_bc       8828
mths_since_recent_inq     27868
num_accts_ever_120_pd     27741
num_actv_bc_tl            27741
num_actv_rev_tl           27741
num_bc_sats               16055
num_bc_tl                 27741
num_il_tl                 27741
num_op_rev_tl             27741
num_rev_accts             27741
num_rev_tl_bal_gt_0       27741
num_tl_op_past_12m        27741
pct_tl_nvr_dlq            27894
```

```
percent_bc_gt_75                9028
pub_rec_bankruptcies               0
tot_hi_cred_lim                27741
total_bc_limit                  7495
total_il_high_credit_limit     27741
Length: 65, dtype: int64
```

In [9]:

```
# df = df.drop(df[df['revol_util']==0].index)
# df.drop(df.index[df['last_credit_pull_d'] == 0], inplace = Tru
e)
```

In [9]:

```
df['int_rate'] = df['int_rate'].str.rstrip('%')
df['int_rate'] = df['int_rate'].astype('float64')
```

In [10]:

```
numericalList = []
nonNumList = []
for column in df.columns:
    if df[column].dtypes == 'int64' or df[column].dtypes == 'flo
at64':
        numericalList.append(column)
    else:
        nonNumList.append(column)
```

In [ ]:

```

```

In [11]:

```
numDF = df[numericalList]
nonnumDF = df[nonNumList]
```

In [12]:

```
nrow = len(numDF['int_rate'])
# print('Number of missing values per column:')
num_countMissing = numDF.isnull().sum()
# print(num_countMissing)
```

In [13]:

```
# Estimate the mean for columns that are missing less than 20% o
f the observations
for coln in numDF.columns:
    if num_countMissing[coln] != 0:
        if num_countMissing[coln]/nrow < 0.20:
            numDF[coln] = numDF[coln].fillna(value=numDF[coln].m
ean())
```

```
/Users/michellebaginski/anaconda3/lib/python3.7/site
-packages/ipykernel_launcher.py:5: SettingWithCopyWa
rning:
A value is trying to be set on a copy of a slice fro
m a DataFrame.
Try using .loc[row_indexer,col_indexer] = value inst
ead

See the caveats in the documentation: http://pandas.
pydata.org/pandas-docs/stable/indexing.html#indexing
-view-versus-copy
  """
```

In [14]:

```
# remove all the observations with NAs in them
df = df.dropna(axis=0,how='any')

print('Number of missing values per column:')
countMissing = df.isnull().sum()
print(countMissing)
```

```
Number of missing values per column:
loan_amnt                       0
funded_amnt                     0
funded_amnt_inv                 0
term                            0
```

```
int_rate                    0
installment                 0
grade                       0
sub_grade                   0
home_ownership              0
annual_inc                  0
verification_status         0
issue_d                     0
loan_status                 0
pymnt_plan                  0
desc                        0
purpose                     0
dti                         0
delinq_2yrs                 0
earliest_cr_line            0
inq_last_6mths              0
open_acc                    0
pub_rec                     0
revol_bal                   0
revol_util                  0
total_acc                   0
initial_list_status         0
total_pymnt                 0
total_pymnt_inv             0
total_rec_prncp             0
total_rec_int               0
                           ..
last_credit_pull_d          0
tot_coll_amt                0
total_rev_hi_lim            0
acc_open_past_24mths        0
avg_cur_bal                 0
bc_open_to_buy              0
bc_util                     0
mo_sin_old_il_acct          0
mo_sin_old_rev_tl_op        0
mo_sin_rcnt_rev_tl_op       0
mo_sin_rcnt_tl              0
mort_acc                    0
mths_since_recent_bc        0
mths_since_recent_inq       0
num_accts_ever_120_pd       0
num_actv_bc_tl              0
num_actv_rev_tl             0
num_bc_sats                 0
```

```
num_bc_tl                     0
num_il_tl                     0
num_op_rev_tl                 0
num_rev_accts                 0
num_rev_tl_bal_gt_0           0
num_tl_op_past_12m            0
pct_tl_nvr_dlq                0
percent_bc_gt_75              0
pub_rec_bankruptcies          0
tot_hi_cred_lim               0
total_bc_limit                0
total_il_high_credit_limit    0
Length: 65, dtype: int64
```

In [15]:

```
# Removing variables on the basis of correlation matrix
corr = numDF.corr()
print(corr)
columns = np.full((corr.shape[0],), True, dtype=bool)
for i in range(corr.shape[0]):
    for j in range(i+1, corr.shape[0]):
        if corr.iloc[i,j] >= 0.8:
            if columns[j]:
                columns[j] = False
selected_columns = numDF.columns[columns]
```

```
                          loan_amnt  funded_amnt
funded_amnt_inv  int_rate  \
loan_amnt                  1.000000     0.999799
0.999663  0.182654
funded_amnt                0.999799     1.000000
0.999874  0.182485
funded_amnt_inv            0.999663     0.999874
1.000000  0.182933
int_rate                   0.182654     0.182485
0.182933  1.000000
installment                0.955011     0.955254
0.955211  0.165173
annual_inc                 0.368164     0.368151
0.368084 -0.026026
dti                        0.044557     0.044572
0.044746  0.147471
delinq_2yrs                0.011184     0.011214
0.011391  0.097230
```

```
inq_last_6mths              0.019741     0.019703
0.020091  0.241345
open_acc                    0.191571     0.191614
0.191719  0.017359
pub_rec                    -0.073588    -0.073537
-0.073373  0.056575
revol_bal                   0.320262     0.320284
0.320259 -0.003159
total_acc                   0.238362     0.238363
0.238460 -0.019417
total_pymnt                 0.891293     0.891500
0.891614  0.203856
total_pymnt_inv             0.891239     0.891456
0.891635  0.204200
total_rec_prncp             0.844267     0.844501
0.844561  0.041107
total_rec_int               0.696224     0.696306
0.696515  0.480275
total_rec_late_fee          0.079565     0.079627
0.079690  0.079482
recoveries                  0.190052     0.190055
0.190051  0.181272
collection_recovery_fee     0.156731     0.156767
0.156717  0.139377
last_pymnt_amnt             0.432104     0.432127
0.432178  0.126608
tot_coll_amt               -0.019016    -0.019020
-0.019030  0.010068
total_rev_hi_lim            0.237312     0.237367
0.237346 -0.143858
acc_open_past_24mths        0.003601     0.003557
0.003667  0.150894
avg_cur_bal                 0.216832     0.216889
0.216840 -0.127608
bc_open_to_buy              0.175445     0.175497
0.175417 -0.340119
bc_util                     0.040698     0.040695
0.040768  0.374959
mo_sin_old_il_acct          0.135317     0.135350
0.135358 -0.031711
mo_sin_old_rev_tl_op        0.173151     0.173190
0.173189 -0.109867
mo_sin_rcnt_rev_tl_op       0.043687     0.043701
0.043624 -0.100733
mo_sin_rcnt_tl              0.008197     0.008203
```

```
0.008097 -0.124643
mort_acc                      0.234556     0.234613
0.234613 -0.096496
mths_since_recent_bc          0.035696     0.035692
0.035657 -0.047071
mths_since_recent_inq        -0.000066    -0.000041
-0.000482 -0.209414
num_accts_ever_120_pd        -0.049013    -0.049041
-0.049017  0.073939
num_actv_bc_tl                0.153285     0.153323
0.153374  0.034850
num_actv_rev_tl               0.125833     0.125862
0.125937  0.124939
num_bc_sats                   0.183156     0.183154
0.183168 -0.061437
num_bc_tl                     0.173745     0.173784
0.173779 -0.092251
num_il_tl                     0.089742     0.089753
0.089812  0.038724
num_op_rev_tl                 0.152853     0.152881
0.152924 -0.009553
num_rev_accts                 0.172846     0.172877
0.172889 -0.060861
num_rev_tl_bal_gt_0           0.125901     0.125930
0.126005  0.125288
num_tl_op_past_12m           -0.008921    -0.008948
-0.008826  0.184984
pct_tl_nvr_dlq                0.072673     0.072692
0.072627 -0.115032
percent_bc_gt_75              0.007203     0.007233
0.007298  0.353748
pub_rec_bankruptcies         -0.094852    -0.094798
-0.094606  0.048524
tot_hi_cred_lim               0.306396     0.306465
0.306440 -0.155279
total_bc_limit                0.358044     0.358130
0.358048 -0.261619
total_il_high_credit_limit    0.173811     0.173821
0.173894  0.024808

                           installment  annual_inc
dti  delinq_2yrs  \
loan_amnt                    0.955011    0.368164
0.044557     0.011184
```

```
funded_amnt                  0.955254    0.368151
0.044572     0.011214
funded_amnt_inv              0.955211    0.368084
0.044746     0.011391
int_rate                     0.165173   -0.026026
0.147471     0.097230
installment                  1.000000    0.367857
0.039438     0.022610
annual_inc                   0.367857    1.000000
-0.196529     0.069232
dti                          0.039438   -0.196529
1.000000    -0.009784
delinq_2yrs                  0.022610    0.069232
-0.009784     1.000000
inq_last_6mths               0.039678    0.085247
0.011601     0.025841
open_acc                     0.187049    0.159963
0.302366     0.056357
pub_rec                     -0.065317   -0.023327
-0.053917    -0.022420
revol_bal                    0.310904    0.341650
0.145847    -0.024285
total_acc                    0.221266    0.238645
0.230711     0.134704
total_pymnt                  0.838972    0.336200
0.042459     0.020337
total_pymnt_inv              0.839014    0.336202
0.042551     0.020470
total_rec_prncp              0.821796    0.347716
0.003891     0.005934
total_rec_int                0.596202    0.203455
0.108085     0.044175
total_rec_late_fee           0.078131    0.032474
0.011868     0.030405
recoveries                   0.162123    0.034279
0.049467     0.015664
collection_recovery_fee      0.129182    0.030331
0.046153     0.015142
last_pymnt_amnt              0.385657    0.191088
-0.028060     0.001715
tot_coll_amt                -0.017793   -0.001675
-0.013457     0.004088
total_rev_hi_lim             0.217850    0.263770
0.048801    -0.023461
acc_open_past_24mths         0.013906    0.049138
```

0.157777 -0.058441
avg_cur_bal 0.186223 0.374005
-0.119971 0.063175
bc_open_to_buy 0.138873 0.166127
-0.091126 -0.032432
bc_util 0.073825 -0.019572
0.206253 -0.017904
mo_sin_old_il_acct 0.117916 0.134886
0.038983 0.079093
mo_sin_old_rev_tl_op 0.151669 0.150621
0.035009 0.094795
mo_sin_rcnt_rev_tl_op 0.029200 0.038323
-0.023834 0.039344
mo_sin_rcnt_tl -0.001376 -0.028659
-0.093251 0.023639
mort_acc 0.197396 0.274060
-0.042512 0.105114
mths_since_recent_bc 0.024574 0.043251
-0.002119 0.069526
mths_since_recent_inq -0.018969 -0.050274
0.000968 -0.018265
num_accts_ever_120_pd -0.038043 0.028157
-0.057088 0.216533
num_actv_bc_tl 0.161251 0.074031
0.147802 -0.058097
num_actv_rev_tl 0.139411 0.046627
0.231110 -0.025010
num_bc_sats 0.181600 0.101672
0.093352 -0.047838
num_bc_tl 0.167666 0.131620
0.064316 0.041219
num_il_tl 0.080364 0.131219
0.239445 0.083041
num_op_rev_tl 0.153466 0.068402
0.157211 0.006416
num_rev_accts 0.166111 0.120787
0.114310 0.082627
num_rev_tl_bal_gt_0 0.139449 0.046725
0.231599 -0.024370
num_tl_op_past_12m 0.008082 0.050567
0.097471 -0.036260
pct_tl_nvr_dlq 0.053564 -0.023768
0.083720 -0.436693
percent_bc_gt_75 0.036409 -0.038702

```
0.188117    -0.021144
pub_rec_bankruptcies          -0.088497   -0.054342
-0.054327    -0.038079
tot_hi_cred_lim                0.270641    0.481292
-0.004814     0.080387
total_bc_limit                 0.323503    0.288585
0.031969    -0.059412
total_il_high_credit_limit     0.164982    0.292326
0.322581     0.068556

                            inq_last_6mths  open_acc
...  num_op_rev_tl  \
loan_amnt                         0.019741  0.191571
...       0.152853
funded_amnt                       0.019703  0.191614
...       0.152881
funded_amnt_inv                   0.020091  0.191719
...       0.152924
int_rate                          0.241345  0.017359
...      -0.009553
installment                       0.039678  0.187049
...       0.153466
annual_inc                        0.085247  0.159963
...       0.068402
dti                               0.011601  0.302366
...       0.157211
delinq_2yrs                       0.025841  0.056357
...       0.006416
inq_last_6mths                    1.000000  0.125784
...       0.089312
open_acc                          0.125784  1.000000
...       0.760595
pub_rec                           0.010963 -0.033834
...      -0.018137
revol_bal                         0.008154  0.217770
...       0.214840
total_acc                         0.154447  0.666391
...       0.459027
total_pymnt                       0.007352  0.167302
...       0.131121
total_pymnt_inv                   0.007723  0.167393
...       0.131154
total_rec_prncp                  -0.016885  0.159681
...       0.127439
total_rec_int                     0.051162  0.125943
```

```
...       0.093824
total_rec_late_fee                0.013419  0.010754
...      -0.002819
recoveries                        0.043446  0.043598
...       0.032373
collection_recovery_fee           0.026198  0.041811
...       0.032996
last_pymnt_amnt                   0.055317  0.082393
...       0.048531
tot_coll_amt                      0.011059  0.005173
...       0.002767
total_rev_hi_lim                  0.030182  0.245941
...       0.299210
acc_open_past_24mths              0.219278  0.436268
...       0.328295
avg_cur_bal                       0.048265 -0.081989
...      -0.190952
bc_open_to_buy                    0.037227  0.237378
...       0.280663
bc_util                          -0.081087 -0.086000
...      -0.124842
mo_sin_old_il_acct                0.013917  0.110979
...       0.055859
mo_sin_old_rev_tl_op             -0.002869  0.132323
...       0.193083
mo_sin_rcnt_rev_tl_op            -0.137293 -0.211942
...      -0.277887
mo_sin_rcnt_tl                   -0.202117 -0.212953
...      -0.186244
mort_acc                          0.097495  0.128045
...       0.056109
mths_since_recent_bc             -0.091008 -0.187125
...      -0.223272
mths_since_recent_inq            -0.640108 -0.085326
...      -0.064062
num_accts_ever_120_pd             0.048480  0.007914
...      -0.019155
num_actv_bc_tl                    0.016336  0.471667
...       0.636410
num_actv_rev_tl                   0.047663  0.597288
...       0.794731
num_bc_sats                       0.052308  0.582938
...       0.725595
num_bc_tl                         0.086334  0.430927
```

```
...       0.563712
num_il_tl                          0.089378  0.354235
...      -0.008681
num_op_rev_tl                      0.089312  0.760595
...       1.000000
num_rev_accts                      0.109858  0.568211
...       0.730405
num_rev_tl_bal_gt_0                0.048051  0.598248
...       0.795936
num_tl_op_past_12m                 0.244767  0.297025
...       0.257699
pct_tl_nvr_dlq                    -0.017770  0.077882
...       0.106137
percent_bc_gt_75                  -0.078196 -0.092029
...      -0.128804
pub_rec_bankruptcies               0.004385 -0.048503
...      -0.026013
tot_hi_cred_lim                    0.097522  0.255636
...       0.111189
total_bc_limit                     0.009950  0.301572
...       0.340577
total_il_high_credit_limit         0.094896  0.334169
...       0.002144

                           num_rev_accts  num_rev_t
l_bal_gt_0  \
loan_amnt                       0.172846
0.125901
funded_amnt                     0.172877
0.125930
funded_amnt_inv                 0.172889
0.126005
int_rate                       -0.060861
0.125288
installment                     0.166111
0.139449
annual_inc                      0.120787
0.046725
dti                             0.114310
0.231599
delinq_2yrs                     0.082627
-0.024370
inq_last_6mths                  0.109858
0.048051
open_acc                        0.568211
```

0.598248
pub_rec -0.012743
-0.017076
revol_bal 0.202645
0.235843
total_acc 0.706895
0.306916
total_pymnt 0.144342
0.120143
total_pymnt_inv 0.144347
0.120207
total_rec_prncp 0.153070
0.092193
total_rec_int 0.075496
0.139935
total_rec_late_fee -0.005876
0.007195
recoveries 0.027268
0.043519
collection_recovery_fee 0.028624
0.042869
last_pymnt_amnt 0.095749
0.007208
tot_coll_amt 0.022182
-0.013133
total_rev_hi_lim 0.271473
0.197993
acc_open_past_24mths 0.283202
0.231467
avg_cur_bal -0.045717
-0.182119
bc_open_to_buy 0.260654
0.061764
bc_util -0.132351
0.121878
mo_sin_old_il_acct 0.151079
0.055949
mo_sin_old_rev_tl_op 0.340432
0.136856
mo_sin_rcnt_rev_tl_op -0.224745
-0.217148
mo_sin_rcnt_tl -0.162482
-0.136792
mort_acc 0.204279

```
0.015141
mths_since_recent_bc            -0.167579
-0.178550
mths_since_recent_inq           -0.078780
-0.037753
num_accts_ever_120_pd            0.107531
-0.029712
num_actv_bc_tl                   0.392016
0.790457
num_actv_rev_tl                  0.509936
0.998491
num_bc_sats                      0.498947
0.628752
num_bc_tl                        0.850541
0.416351
num_il_tl                        0.081672
-0.021621
num_op_rev_tl                    0.730405
0.795936
num_rev_accts                    1.000000
0.510600
num_rev_tl_bal_gt_0              0.510600
1.000000
num_tl_op_past_12m               0.226864
0.178580
pct_tl_nvr_dlq                   0.004681
0.097005
percent_bc_gt_75                -0.132262
0.098695
pub_rec_bankruptcies            -0.014779
-0.023540
tot_hi_cred_lim                  0.180928
0.059625
total_bc_limit                   0.313148
0.196284
total_il_high_credit_limit       0.043024
-0.012111

                          num_tl_op_past_12m  pct_
tl_nvr_dlq  \
loan_amnt                         -0.008921
0.072673
funded_amnt                       -0.008948
0.072692
funded_amnt_inv                   -0.008826
```

0.072627
int_rate 0.184984
-0.115032
installment 0.008082
0.053564
annual_inc 0.050567
-0.023768
dti 0.097471
0.083720
delinq_2yrs -0.036260
-0.436693
inq_last_6mths 0.244767
-0.017770
open_acc 0.297025
0.077882
pub_rec 0.000880
0.006628
revol_bal -0.024168
0.116028
total_acc 0.262375
-0.016746
total_pymnt -0.026924
0.049632
total_pymnt_inv -0.026818
0.049573
total_rec_prncp -0.046609
0.063754
total_rec_int 0.014470
0.001775
total_rec_late_fee 0.009647
-0.023673
recoveries 0.050521
0.002297
collection_recovery_fee 0.035131
0.002102
last_pymnt_amnt 0.044297
0.040964
tot_coll_amt 0.011436
-0.062637
total_rev_hi_lim 0.036456
0.125221
acc_open_past_24mths 0.665080
0.047643
avg_cur_bal -0.018951

-0.040238
bc_open_to_buy 0.063405
0.121734
bc_util -0.130236
-0.001699
mo_sin_old_il_acct -0.004021
-0.096722
mo_sin_old_rev_tl_op -0.024198
-0.106522
mo_sin_rcnt_rev_tl_op -0.433578
-0.037148
mo_sin_rcnt_tl -0.534940
-0.030447
mort_acc 0.067999
-0.045008
mths_since_recent_bc -0.287923
-0.054641
mths_since_recent_inq -0.225449
0.017137
num_accts_ever_120_pd 0.057674
-0.550597
num_actv_bc_tl 0.083757
0.127743
num_actv_rev_tl 0.177864
0.096877
num_bc_sats 0.140359
0.137486
num_bc_tl 0.155990
0.018570
num_il_tl 0.196592
-0.014351
num_op_rev_tl 0.257699
0.106137
num_rev_accts 0.226864
0.004681
num_rev_tl_bal_gt_0 0.178580
0.097005
num_tl_op_past_12m 1.000000
0.015170
pct_tl_nvr_dlq 0.015170
1.000000
percent_bc_gt_75 -0.123374
0.009183
pub_rec_bankruptcies -0.002445
0.033255

```
tot_hi_cred_lim                      0.100894
0.008941
total_bc_limit                       0.004563
0.191606
total_il_high_credit_limit           0.151449
-0.009095

                           percent_bc_gt_75  pub_re
c_bankruptcies  \
loan_amnt                          0.007203
-0.094852
funded_amnt                        0.007233
-0.094798
funded_amnt_inv                    0.007298
-0.094606
int_rate                           0.353748
0.048524
installment                        0.036409
-0.088497
annual_inc                        -0.038702
-0.054342
dti                                0.188117
-0.054327
delinq_2yrs                       -0.021144
-0.038079
inq_last_6mths                    -0.078196
0.004385
open_acc                          -0.092029
-0.048503
pub_rec                           -0.025297
0.759816
revol_bal                          0.087455
-0.105968
total_acc                         -0.079553
-0.020924
total_pymnt                        0.029349
-0.083643
total_pymnt_inv                    0.029401
-0.083494
total_rec_prncp                   -0.024151
-0.086063
total_rec_int                      0.135720
-0.050333
total_rec_late_fee                 0.026717
```

-0.015232
recoveries 0.040779
-0.013886
collection_recovery_fee 0.033718
-0.008255
last_pymnt_amnt -0.026677
-0.021988
tot_coll_amt -0.028502
0.018787
total_rev_hi_lim -0.137056
-0.093797
acc_open_past_24mths -0.116840
0.011326
avg_cur_bal 0.016887
-0.060993
bc_open_to_buy -0.477066
-0.087191
bc_util 0.831412
-0.012362
mo_sin_old_il_acct 0.029971
0.042991
mo_sin_old_rev_tl_op -0.026080
0.035970
mo_sin_rcnt_rev_tl_op 0.086981
-0.033634
mo_sin_rcnt_tl 0.081118
-0.012071
mort_acc -0.039093
0.000999
mths_since_recent_bc 0.123060
-0.009216
mths_since_recent_inq 0.055103
-0.006782
num_accts_ever_120_pd -0.027612
-0.004829
num_actv_bc_tl 0.040894
-0.037957
num_actv_rev_tl 0.098692
-0.023378
num_bc_sats -0.177885
-0.046185
num_bc_tl -0.154246
-0.015137
num_il_tl 0.027746
-0.023903

```
num_op_rev_tl                        -0.128804
-0.026013
num_rev_accts                        -0.132262
-0.014779
num_rev_tl_bal_gt_0                   0.098695
-0.023540
num_tl_op_past_12m                   -0.123374
-0.002445
pct_tl_nvr_dlq                        0.009183
0.033255
percent_bc_gt_75                      1.000000
-0.016911
pub_rec_bankruptcies                 -0.016911
1.000000
tot_hi_cred_lim                      -0.048225
-0.088544
total_bc_limit                       -0.249861
-0.142249
total_il_high_credit_limit            0.010729
-0.039593

                          tot_hi_cred_lim  total_b
c_limit  \
loan_amnt                        0.306396        0
.358044
funded_amnt                      0.306465        0
.358130
funded_amnt_inv                  0.306440        0
.358048
int_rate                        -0.155279       -0
.261619
installment                      0.270641        0
.323503
annual_inc                       0.481292        0
.288585
dti                             -0.004814        0
.031969
delinq_2yrs                      0.080387       -0
.059412
inq_last_6mths                   0.097522        0
.009950
open_acc                         0.255636        0
.301572
pub_rec                         -0.063181       -0
```

.116149
revol_bal 0.448612 0
.478627
total_acc 0.327509 0
.257766
total_pymnt 0.274987 0
.309217
total_pymnt_inv 0.274958 0
.309131
total_rec_prncp 0.296321 0
.348431
total_rec_int 0.140758 0
.125307
total_rec_late_fee 0.018769 -0
.008083
recoveries 0.019262 0
.011394
collection_recovery_fee 0.023102 0
.012161
last_pymnt_amnt 0.169447 0
.155508
tot_coll_amt -0.002933 -0
.031221
total_rev_hi_lim 0.447925 0
.561239
acc_open_past_24mths 0.101169 -0
.009808
avg_cur_bal 0.822000 0
.151672
bc_open_to_buy 0.233889 0
.839838
bc_util -0.045084 -0
.290919
mo_sin_old_il_acct 0.181827 0
.104874
mo_sin_old_rev_tl_op 0.205132 0
.265807
mo_sin_rcnt_rev_tl_op 0.027206 0
.007140
mo_sin_rcnt_tl -0.075786 0
.010811
mort_acc 0.512527 0
.220792
mths_since_recent_bc 0.040567 -0
.062443

```
mths_since_recent_inq              -0.059908        0
.003645
num_accts_ever_120_pd               0.006180       -0
.123785
num_actv_bc_tl                      0.060523        0
.369297
num_actv_rev_tl                     0.059402        0
.195929
num_bc_sats                         0.103953        0
.520473
num_bc_tl                           0.153889        0
.396479
num_il_tl                           0.201089       -0
.016023
num_op_rev_tl                       0.111189        0
.340577
num_rev_accts                       0.180928        0
.313148
num_rev_tl_bal_gt_0                 0.059625        0
.196284
num_tl_op_past_12m                  0.100894        0
.004563
pct_tl_nvr_dlq                      0.008941        0
.191606
percent_bc_gt_75                   -0.048225       -0
.249861
pub_rec_bankruptcies               -0.088544       -0
.142249
tot_hi_cred_lim                     1.000000        0
.353012
total_bc_limit                      0.353012        1
.000000
total_il_high_credit_limit          0.375659        0
.072251

                           total_il_high_credit_lim
it
loan_amnt                                  0.1738
11
funded_amnt                                0.1738
21
funded_amnt_inv                            0.1738
94
int_rate                                   0.0248
```

08
installment 0.1649
82
annual_inc 0.2923
26
dti 0.3225
81
delinq_2yrs 0.0685
56
inq_last_6mths 0.0948
96
open_acc 0.3341
69
pub_rec -0.0238
02
revol_bal 0.0944
52
total_acc 0.3821
61
total_pymnt 0.1585
47
total_pymnt_inv 0.1586
09
total_rec_prncp 0.1516
23
total_rec_int 0.1204
54
total_rec_late_fee 0.0250
26
recoveries 0.0338
33
collection_recovery_fee 0.0329
16
last_pymnt_amnt 0.0978
72
tot_coll_amt 0.0038
19
total_rev_hi_lim 0.0687
52
acc_open_past_24mths 0.2006
55
avg_cur_bal 0.1883
51
bc_open_to_buy 0.0272
64

bc_util 0.0194
64
mo_sin_old_il_acct 0.1794
61
mo_sin_old_rev_tl_op 0.0187
94
mo_sin_rcnt_rev_tl_op 0.0051
09
mo_sin_rcnt_tl -0.1201
61
mort_acc 0.1050
41
mths_since_recent_bc 0.0072
51
mths_since_recent_inq -0.0526
63
num_accts_ever_120_pd 0.0417
45
num_actv_bc_tl -0.0101
72
num_actv_rev_tl -0.0122
39
num_bc_sats 0.0028
15
num_bc_tl 0.0331
86
num_il_tl 0.6056
54
num_op_rev_tl 0.0021
44
num_rev_accts 0.0430
24
num_rev_tl_bal_gt_0 -0.0121
11
num_tl_op_past_12m 0.1514
49
pct_tl_nvr_dlq -0.0090
95
percent_bc_gt_75 0.0107
29
pub_rec_bankruptcies -0.0395
93
tot_hi_cred_lim 0.3756
59

```
total_bc_limit                                0.0722
51
total_il_high_credit_limit                    1.0000
00

[50 rows x 50 columns]
```

In [17]:

```
numDF = numDF[selected_columns]
```

In [18]:

```
# £print(nonnumDF[['grade','home_ownership','loan_status','purpo
se','sub_grade','term']])
```

In [19]:

```
# Quartiles and IQR for mths_since_recent_bc

for col in numDF.columns:
    quartiles = numDF[col].quantile([0.25, 0.75], interpolation=
'nearest')
    q1 = quartiles[0.25]
    q3 = quartiles[0.75]
    IQR = q3 - q1
    outlier_val = q3 + 1.5*IQR
    #print("Outlier val:", outlier_val)
    numDF[col] = np.where(numDF[col] > outlier_val, outlier_val,
numDF[col])

#df1 = df['mths_since_recent_bc'] = np.where(df['mths_since_rece
nt_bc'] > outlier_val, outlier_val, df['mths_since_recent_bc'])
```

In [17]:

```
finaldf = pd.concat([numDF, nonnumDF[['grade','home_ownership','
loan_status','purpose','sub_grade','term']]], axis=1)
```

In [18]:

```
print('Number of missing values per column:')
countMissing = finaldf['loan_status']
count = 0
ncount = 0
for w in countMissing:
    if w == "Fully Paid":
        count+=1
    else:
        ncount+=1
print(count/len(finaldf['loan_status']))
print(ncount/len(finaldf['loan_status']))
```

```
Number of missing values per column:
0.8423273337903402
0.15767266620965986
```

In [19]:

```
#making categorical variable dummy variable
finaldf = pd.get_dummies(finaldf, columns=['grade','home_ownersh
ip','loan_status','purpose','sub_grade','term'])
```

In [20]:

```
print(finaldf.columns[100:])
```

```
Index(['sub_grade_E4', 'sub_grade_E5', 'sub_grade_F1
', 'sub_grade_F2',
       'sub_grade_F3', 'sub_grade_F4', 'sub_grade_F5
', 'sub_grade_G1',
       'sub_grade_G2', 'sub_grade_G3', 'sub_grade_G4
', 'sub_grade_G5',
       'term_ 36 months', 'term_ 60 months'],
      dtype='object')
```

In [21]:

```
## Dividing dataset into 2 parts
x_train, x_test, y_train, y_test = train_test_split(finaldf.loc[
:, finaldf.columns != 'int_rate'], finaldf['int_rate'], test_siz
e=0.3, random_state = 0)
```

In [22]:

```
scaler = StandardScaler()
scaler.fit(x_train)
scaler.fit(x_test)
names = x_train.columns
x_train_scaled = scaler.transform(x_train)
x_train_scaled = pd.DataFrame(x_train_scaled, columns=names)
x_test_scaled = scaler.transform(x_test)
x_test_scaled = pd.DataFrame(x_test_scaled, columns=names)
```

In [23]:

```
# Build model with only numeric variables (using statsmodels)
#df=df.rename(columns = {"Percent Less than Bachelors Degree":Pe
rcent Less than Bachelor Degree})
model = smf.ols(formula = 'int_rate ~ loan_amnt + annual_inc + d
ti + delinq_2yrs + inq_last_6mths+open_acc + pub_rec + revol_bal
+ total_acc + total_rec_late_fee+ recoveries + last_pymnt_amnt +
tot_coll_amt + acc_open_past_24mths + avg_cur_bal + bc_open_to_b
uy +bc_util + mo_sin_old_il_acct + mo_sin_old_rev_tl_op + mo_sin
_rcnt_rev_tl_op + mo_sin_rcnt_tl + mort_acc + mths_since_recent_
bc + mths_since_recent_inq + num_accts_ever_120_pd + num_actv_bc
_tl + num_actv_rev_tl + num_bc_tl + num_il_tl + num_tl_op_past_1
2m + pct_tl_nvr_dlq  + total_il_high_credit_limit', data = final
df)
results = model.fit()
print(results.summary())
```

```
                            OLS Regression Results
==============================================================
==========================
Dep. Variable:                   int_rate   R-squared:
0.411
Model:                            OLS   Adj. R-squar
ed:                  0.411
Method:                 Least Squares   F-statistic:
```

```
4103.
Date:                Mon, 02 Dec 2019   Prob (F-stat
istic):               0.00
Time:                        15:50:07   Log-Likeliho
od:            -4.9761e+05
No. Observations:             188181   AIC:
9.953e+05
Df Residuals:                 188148   BIC:
9.956e+05
Df Model:                         32
Covariance Type:            nonrobust
================================================
=========================================
                                coef    std err
t      P>|t|      [0.025      0.975]
------------------------------------------------
----------------------------------------
Intercept                     15.2796      0.157
97.379      0.000      14.972      15.587
loan_amnt                      0.0001   1.25e-06
98.418      0.000       0.000       0.000
annual_inc                 -1.84e-06   1.95e-07
-9.420      0.000   -2.22e-06   -1.46e-06
dti                            0.0298      0.001
23.222      0.000       0.027       0.032
delinq_2yrs                    0.5857      0.013
46.173      0.000       0.561       0.611
inq_last_6mths                 0.7765      0.010
76.788      0.000       0.757       0.796
open_acc                      -0.0755      0.003
-24.107      0.000      -0.082      -0.069
pub_rec                        0.7662      0.020
39.029      0.000       0.728       0.805
revol_bal                   1.292e-06   4.94e-07
2.617      0.009    3.24e-07    2.26e-06
total_acc                     -0.0070      0.002
-3.658      0.000      -0.011      -0.003
total_rec_late_fee             0.0215      0.001
21.349      0.000       0.020       0.024
recoveries                     0.0006   9.65e-06
62.001      0.000       0.001       0.001
last_pymnt_amnt             7.961e-05   1.56e-06
51.042      0.000    7.66e-05    8.27e-05
tot_coll_amt                5.745e-05   9.98e-06
5.756      0.000    3.79e-05     7.7e-05
```

```
acc_open_past_24mths           0.1088      0.005
23.913      0.000       0.100       0.118
avg_cur_bal                -3.857e-05   6.55e-07
-58.900      0.000   -3.99e-05   -3.73e-05
bc_open_to_buy             -4.477e-05   8.68e-07
-51.554      0.000   -4.65e-05   -4.31e-05
bc_util                        0.0514      0.000
121.769      0.000       0.051       0.052
mo_sin_old_il_acct            -0.0020      0.000
-10.521      0.000      -0.002      -0.002
mo_sin_old_rev_tl_op          -0.0035      0.000
-32.054      0.000      -0.004      -0.003
mo_sin_rcnt_rev_tl_op          0.0046      0.001
5.728      0.000       0.003       0.006
mo_sin_rcnt_tl                -0.0089      0.001
-7.045      0.000      -0.011      -0.006
mort_acc                      -0.1508      0.005
-30.466      0.000      -0.160      -0.141
mths_since_recent_bc          -0.0025      0.000
-7.149      0.000      -0.003      -0.002
mths_since_recent_inq         -0.0619      0.002
-32.650      0.000      -0.066      -0.058
num_accts_ever_120_pd          0.1198      0.011
10.891      0.000       0.098       0.141
num_actv_bc_tl                -0.0873      0.008
-10.734      0.000      -0.103      -0.071
num_actv_rev_tl                0.1780      0.006
30.415      0.000       0.166       0.189
num_bc_tl                     -0.0701      0.003
-21.982      0.000      -0.076      -0.064
num_il_tl                     -0.0070      0.002
-2.831      0.005      -0.012      -0.002
num_tl_op_past_12m             0.4447      0.008
54.304      0.000       0.429       0.461
pct_tl_nvr_dlq                -0.0575      0.002
-37.437      0.000      -0.060      -0.054
total_il_high_credit_limit -3.497e-06   3.33e-07
-10.510      0.000   -4.15e-06   -2.85e-06
==============================================
==========================
Omnibus:                       4803.670   Durbin-Watso
n:                    1.947
Prob(Omnibus):                    0.000   Jarque-Bera
(JB):              5550.412
```

```
Skew:                          0.355   Prob(JB):
0.00
Kurtosis:                      3.451   Cond. No.
2.00e+06
==============================================================
==========================

Warnings:
[1] Standard Errors assume that the covariance matri
x of the errors is correctly specified.
[2] The condition number is large,  2e+06. This migh
t indicate that there are
strong multicollinearity or other numerical problems
.
```

In [24]:

```
# Build model with full dataset and removing varibales on the ba
(using statsmodels)
#df=df.rename(columns = {"Percent Less than Bachelors Degree":Pe
rcent Less than Bachelor Degree})
model = smf.ols(formula = 'int_rate ~ loan_amnt + annual_inc + d
ti + delinq_2yrs + inq_last_6mths+open_acc + pub_rec + revol_bal
+ total_acc + total_rec_late_fee+ recoveries + last_pymnt_amnt +
tot_coll_amt + acc_open_past_24mths + avg_cur_bal + bc_open_to_b
uy +bc_util + mo_sin_old_il_acct + mo_sin_old_rev_tl_op + mo_sin
_rcnt_rev_tl_op + mo_sin_rcnt_tl + mort_acc + mths_since_recent_
bc + mths_since_recent_inq + num_accts_ever_120_pd + num_actv_bc
_tl + num_actv_rev_tl + num_bc_tl + num_il_tl + num_tl_op_past_1
2m + pct_tl_nvr_dlq  + total_il_high_credit_limit + grade_A+ gra
de_B+ grade_C+ grade_D+ grade_E+ grade_F+ grade_G+ home_ownershi
p_MORTGAGE \
                  + home_ownership_NONE+ home_ownership_OTHER+ hom
e_ownership_OWN+ \
       home_ownership_RENT+ Q("loan_status_Charged Off") \
       + Q("loan_status_Fully Paid") + purpose_car+ purpose_cred
it_card \
        + purpose_debt_consolidation+ purpose_home_improvement \
        + purpose_house+ purpose_major_purchase+ purpose_medical
\
       + purpose_moving+ purpose_other+ purpose_renewable_energy
\
        + purpose_small_business+ purpose_vacation+ purpose_wedd
ing \
       + sub_grade_A1+ sub_grade_A2+ sub_grade_A3+ sub_grade_A4
```

```
\
      + sub_grade_A5+ sub_grade_B1+ sub_grade_B2+ sub_grade_B3
\
      + sub_grade_B4+ sub_grade_B5+ sub_grade_C1+ sub_grade_C2
\
      + sub_grade_C3+ sub_grade_C4+ sub_grade_C5+ sub_grade_D1
\
      + sub_grade_D2+ sub_grade_D3+ sub_grade_D4+ sub_grade_D5
\
      + sub_grade_E1 + sub_grade_E2 + sub_grade_E3 + sub_grade_
E4+ sub_grade_E5+ sub_grade_F1+ sub_grade_F2\
      + sub_grade_F3+ sub_grade_F4+ sub_grade_F5+ sub_grade_G1\
      + sub_grade_G2+ sub_grade_G3+ sub_grade_G4+ sub_grade_G5'
, data = finaldf)
results = model.fit()
print(results.summary())
```

```
                            OLS Regression Results
==============================================================
==========================
Dep. Variable:                   int_rate   R-squared:
0.994
Model:                               OLS   Adj. R-squar
ed:                  0.994
Method:                   Least Squares   F-statistic:
3.634e+05
Date:                  Mon, 02 Dec 2019   Prob (F-stat
istic):                0.00
Time:                          15:50:55   Log-Likeliho
od:                -69087.
No. Observations:                188181   AIC:
1.383e+05
Df Residuals:                    188097   BIC:
1.392e+05
Df Model:                            83
Covariance Type:             nonrobust
==============================================================
==========================================
                                    coef    std err
t      P>|t|      [0.025      0.975]
----------------------------------------------------------
------------------------------------------
Intercept                          9.1528      0.011
803.264      0.000       9.130       9.175
```

```
loan_amnt                   -1.894e-06   1.37e-07
-13.813      0.000   -2.16e-06   -1.62e-06
annual_inc                   8.221e-08   2.01e-08
4.085      0.000    4.28e-08    1.22e-07
dti                             0.0020      0.000
14.874      0.000       0.002       0.002
delinq_2yrs                     0.0058      0.001
4.458      0.000       0.003       0.008
inq_last_6mths                  0.0031      0.001
2.895      0.004       0.001       0.005
open_acc                       -0.0007      0.000
-2.074      0.038      -0.001   -3.67e-05
pub_rec                         0.0009      0.002
0.467      0.640      -0.003       0.005
revol_bal                  -2.863e-08   5.08e-08
-0.564      0.573   -1.28e-07    7.09e-08
total_acc                       0.0016      0.000
7.920      0.000       0.001       0.002
total_rec_late_fee           8.397e-05      0.000
0.809      0.419      -0.000       0.000
recoveries                   5.516e-06   1.13e-06
4.881      0.000     3.3e-06    7.73e-06
last_pymnt_amnt              1.041e-06   1.67e-07
6.238      0.000    7.14e-07    1.37e-06
tot_coll_amt                 7.904e-07   1.02e-06
0.772      0.440   -1.22e-06     2.8e-06
acc_open_past_24mths           -0.0018      0.000
-3.831      0.000      -0.003      -0.001
avg_cur_bal                 -2.29e-07   7.07e-08
-3.238      0.001   -3.68e-07   -9.04e-08
bc_open_to_buy             -6.738e-08   9.16e-08
-0.736      0.462   -2.47e-07    1.12e-07
bc_util                         0.0005   4.58e-05
11.298      0.000       0.000       0.001
mo_sin_old_il_acct         -1.784e-05   1.98e-05
-0.900      0.368   -5.67e-05     2.1e-05
mo_sin_old_rev_tl_op       -5.725e-05   1.14e-05
-5.018      0.000   -7.96e-05   -3.49e-05
mo_sin_rcnt_rev_tl_op      -1.442e-05    8.2e-05
-0.176      0.860      -0.000       0.000
mo_sin_rcnt_tl                 -0.0001      0.000
-0.962      0.336      -0.000       0.000
mort_acc                       -0.0029      0.001
-5.492      0.000      -0.004      -0.002
mths_since_recent_bc       -4.408e-05   3.55e-05
```

```
-1.243      0.214      -0.000    2.54e-05
mths_since_recent_inq           -0.0011      0.000
-5.766      0.000      -0.002      -0.001
num_accts_ever_120_pd            0.0010      0.001
0.900      0.368      -0.001       0.003
num_actv_bc_tl                   0.0040      0.001
4.791      0.000       0.002       0.006
num_actv_rev_tl                 -0.0011      0.001
-1.908      0.056      -0.002    3.13e-05
num_bc_tl                       -0.0019      0.000
-5.912      0.000      -0.003      -0.001
num_il_tl                       -0.0023      0.000
-9.238      0.000      -0.003      -0.002
num_tl_op_past_12m               0.0088      0.001
10.334      0.000       0.007       0.010
pct_tl_nvr_dlq                  -0.0004      0.000
-2.240      0.025      -0.001   -4.43e-05
total_il_high_credit_limit   -6.149e-08   3.43e-08
-1.794      0.073   -1.29e-07    5.69e-09
grade_A                         -7.3014      0.003
-2301.683      0.000      -7.308      -7.295
grade_B                         -3.5852      0.003
-1383.448      0.000      -3.590      -3.580
grade_C                         -0.5565      0.003
-220.365      0.000      -0.561      -0.552
grade_D                          2.0939      0.003
774.588      0.000       2.089       2.099
grade_E                          4.5038      0.003
1392.196      0.000       4.497       4.510
grade_F                          6.3899      0.004
1519.392      0.000       6.382       6.398
grade_G                          7.6083      0.009
821.128      0.000       7.590       7.626
home_ownership_MORTGAGE          1.7934      0.014
132.113      0.000       1.767       1.820
home_ownership_NONE              1.8838      0.045
41.648      0.000       1.795       1.972
home_ownership_OTHER             1.8860      0.043
43.442      0.000       1.801       1.971
home_ownership_OWN               1.7989      0.014
131.236      0.000       1.772       1.826
home_ownership_RENT              1.7907      0.014
131.985      0.000       1.764       1.817
Q("loan_status_Charged Off")     4.5726      0.006
770.300      0.000       4.561       4.584
```

```
Q("loan_status_Fully Paid")      4.5802      0.006
791.847      0.000       4.569       4.592
purpose_car                      0.6736      0.008
84.346      0.000       0.658       0.689
purpose_credit_card              0.6911      0.004
189.130      0.000       0.684       0.698
purpose_debt_consolidation       0.6945      0.003
203.161      0.000       0.688       0.701
purpose_home_improvement         0.6986      0.005
153.373      0.000       0.690       0.708
purpose_house                    0.7048      0.010
68.459      0.000       0.685       0.725
purpose_major_purchase           0.6928      0.006
111.409      0.000       0.681       0.705
purpose_medical                  0.7309      0.009
82.288      0.000       0.714       0.748
purpose_moving                   0.7461      0.011
70.839      0.000       0.725       0.767
purpose_other                    0.7220      0.005
153.700      0.000       0.713       0.731
purpose_renewable_energy         0.7316      0.029
24.900      0.000       0.674       0.789
purpose_small_business           0.6846      0.007
98.441      0.000       0.671       0.698
purpose_vacation                 0.7035      0.011
62.873      0.000       0.682       0.725
purpose_wedding                  0.6786      0.009
72.214      0.000       0.660       0.697
sub_grade_A1                    -2.8500      0.005
-580.785      0.000      -2.860      -2.840
sub_grade_A2                    -2.2376      0.005
-466.453      0.000      -2.247      -2.228
sub_grade_A3                    -1.2779      0.004
-285.755      0.000      -1.287      -1.269
sub_grade_A4                    -0.9590      0.004
-246.964      0.000      -0.967      -0.951
sub_grade_A5                     0.0230      0.004
6.272      0.000       0.016       0.030
sub_grade_B1                    -2.6896      0.003
-860.035      0.000      -2.696      -2.683
sub_grade_B2                    -1.6498      0.003
-584.933      0.000      -1.655      -1.644
sub_grade_B3                    -0.6404      0.003
-244.339      0.000      -0.646      -0.635
sub_grade_B4                     0.2915      0.003
```

```
108.260      0.000       0.286       0.297
sub_grade_B5                      1.1031      0.003
339.663      0.000       1.097       1.109
sub_grade_C1                     -1.5235      0.003
-500.682      0.000      -1.529      -1.517
sub_grade_C2                     -0.6986      0.003
-225.551      0.000      -0.705      -0.693
sub_grade_C3                     -0.1129      0.003
-35.422      0.000      -0.119      -0.107
sub_grade_C4                      0.4429      0.003
137.732      0.000       0.437       0.449
sub_grade_C5                      1.3356      0.003
397.229      0.000       1.329       1.342
sub_grade_D1                     -0.6970      0.004
-182.936      0.000      -0.704      -0.690
sub_grade_D2                     -0.0162      0.004
-3.940      0.000      -0.024      -0.008
sub_grade_D3                      0.4094      0.004
93.363      0.000       0.401       0.418
sub_grade_D4                      0.8444      0.004
192.044      0.000       0.836       0.853
sub_grade_D5                      1.5533      0.005
330.716      0.000       1.544       1.562
sub_grade_E1                     -0.1881      0.006
-30.101      0.000      -0.200      -0.176
sub_grade_E2                      0.3883      0.006
66.524      0.000       0.377       0.400
sub_grade_E3                      0.9169      0.006
143.103      0.000       0.904       0.929
sub_grade_E4                      1.4574      0.006
224.804      0.000       1.445       1.470
sub_grade_E5                      1.9293      0.007
272.247      0.000       1.915       1.943
sub_grade_F1                      0.5513      0.008
67.368      0.000       0.535       0.567
sub_grade_F2                      0.9400      0.009
107.013      0.000       0.923       0.957
sub_grade_F3                      1.3627      0.009
146.091      0.000       1.344       1.381
sub_grade_F4                      1.6889      0.010
163.044      0.000       1.669       1.709
sub_grade_F5                      1.8470      0.011
161.679      0.000       1.825       1.869
sub_grade_G1                      1.2307      0.017
71.702      0.000       1.197       1.264
```

```
sub_grade_G2                          1.4191      0.020
71.812      0.000        1.380        1.458
sub_grade_G3                          1.6319      0.023
71.798      0.000        1.587        1.676
sub_grade_G4                          1.6392      0.029
56.233      0.000        1.582        1.696
sub_grade_G5                          1.6874      0.032
52.426      0.000        1.624        1.751
==============================================================
==========================
Omnibus:                      275804.446   Durbin-Watso
n:                    1.455
Prob(Omnibus):                     0.000   Jarque-Bera
(JB):         618172593.074
Skew:                             -8.225   Prob(JB):
0.00
Kurtosis:                        283.302   Cond. No.
1.18e+16
==============================================================
==========================

Warnings:
[1] Standard Errors assume that the covariance matri
x of the errors is correctly specified.
[2] The smallest eigenvalue is 1.34e-17. This might
indicate that there are
strong multicollinearity problems or that the design
matrix is singular.
```

In [25]:

```
# Build model after removing variables on the basis of p value(u
sing statsmodels)
#df=df.rename(columns = {"Percent Less than Bachelors Degree":Pe
rcent Less than Bachelor Degree})
model = smf.ols(formula = 'int_rate ~ loan_amnt  + dti  + inq_la
st_6mths + pub_rec +bc_util + num_bc_tl + num_tl_op_past_12m + g
rade_A+ grade_B+ grade_C+ grade_D+ grade_E+ home_ownership_MORTG
AGE \
                + home_ownership_OTHER+ home_ownership_OWN+ \
       home_ownership_RENT+ Q("loan_status_Charged Off") \
        + purpose_credit_card \
        + purpose_debt_consolidation+ purpose_home_improvement \
        + purpose_house+ purpose_major_purchase+ purpose_medical
\
       + purpose_moving+ purpose_other+ purpose_renewable_energy
\
        + purpose_small_business+ purpose_vacation+ purpose_wedd
ing' , data = finaldf)
results = model.fit()
print(results.summary())
```

```
                            OLS Regression Results
==============================================================
==========================
Dep. Variable:                   int_rate   R-squared:
0.943
Model:                            OLS   Adj. R-squar
ed:                  0.943
Method:                 Least Squares   F-statistic:
1.071e+05
Date:                Mon, 02 Dec 2019   Prob (F-stat
istic):                0.00
Time:                        15:51:49   Log-Likeliho
od:            -2.7805e+05
No. Observations:              188181   AIC:
5.562e+05
Df Residuals:                  188151   BIC:
5.565e+05
Df Model:                          29
Covariance Type:            nonrobust
==============================================================
==========================================
```

```
                                    coef    std err
t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept                          22.7685      0.166
136.838      0.000      22.442      23.095
loan_amnt                        1.307e-05   3.37e-07
38.792      0.000    1.24e-05    1.37e-05
dti                                0.0055      0.000
16.329      0.000       0.005       0.006
inq_last_6mths                     0.1032      0.003
40.510      0.000       0.098       0.108
pub_rec                            0.1272      0.006
20.922      0.000       0.115       0.139
bc_util                            0.0083      0.000
73.584      0.000       0.008       0.008
num_bc_tl                         -0.0084      0.001
-14.524      0.000      -0.010      -0.007
num_tl_op_past_12m                 0.0641      0.002
33.927      0.000       0.060       0.068
grade_A                          -15.6427      0.016
-983.692      0.000     -15.674     -15.612
grade_B                          -11.6300      0.015
-797.860      0.000     -11.659     -11.601
grade_C                           -8.2125      0.014
-575.300      0.000      -8.240      -8.184
grade_D                           -5.1964      0.015
-355.634      0.000      -5.225      -5.168
grade_E                           -2.3557      0.016
-146.500      0.000      -2.387      -2.324
home_ownership_MORTGAGE           -0.2606      0.164
-1.592      0.111      -0.581       0.060
home_ownership_OTHER              -0.0634      0.226
-0.280      0.779      -0.507       0.380
home_ownership_OWN                -0.1249      0.164
-0.762      0.446      -0.446       0.196
home_ownership_RENT               -0.0999      0.164
-0.610      0.542      -0.421       0.221
Q("loan_status_Charged Off")       0.1204      0.007
17.512      0.000       0.107       0.134
purpose_credit_card               -0.0170      0.025
-0.686      0.493      -0.065       0.031
purpose_debt_consolidation         0.0530      0.024
2.171      0.030       0.005       0.101
purpose_home_improvement           0.0625      0.026
```

```
2.375      0.018       0.011       0.114
purpose_house                     0.0428      0.040
1.066      0.286      -0.036       0.122
purpose_major_purchase            0.0483      0.030
1.625      0.104      -0.010       0.107
purpose_medical                   0.2844      0.036
7.821      0.000       0.213       0.356
purpose_moving                    0.3000      0.041
7.343      0.000       0.220       0.380
purpose_other                     0.3529      0.027
13.240      0.000       0.301       0.405
purpose_renewable_energy          0.2457      0.099
2.482      0.013       0.052       0.440
purpose_small_business            0.2036      0.032
6.452      0.000       0.142       0.265
purpose_vacation                  0.3329      0.043
7.802      0.000       0.249       0.417
purpose_wedding                   0.1870      0.038
4.951      0.000       0.113       0.261
==============================================================
==========================
Omnibus:                    14152.174   Durbin-Watso
n:                   1.931
Prob(Omnibus):                  0.000   Jarque-Bera
(JB):            55733.726
Skew:                          -0.297   Prob(JB):
0.00
Kurtosis:                       5.599   Cond. No.
2.52e+06
==============================================================
==========================

Warnings:
[1] Standard Errors assume that the covariance matri
x of the errors is correctly specified.
[2] The condition number is large, 2.52e+06. This mi
ght indicate that there are
strong multicollinearity or other numerical problems
.
```

In [26]:

```
#  checking out model to predict the test accuracy
model = linear_model.LinearRegression()
fitted_model = model.fit(X = x_train_scaled[['loan_amnt'  , 'dti
'  , 'inq_last_6mths' , 'pub_rec' , 'bc_util' , 'num_bc_tl' , 'n
um_tl_op_past_12m' , 'grade_A', 'grade_B', 'grade_C', 'grade_D',
'grade_E' , 'home_ownership_MORTGAGE' \
                , 'home_ownership_OTHER' , 'home_ownership_OWN'
, \
       'home_ownership_RENT' , 'loan_status_Charged Off', 'purpo
se_credit_card', 'purpose_debt_consolidation' , 'purpose_home_im
provement', 'purpose_house', 'purpose_major_purchase', 'purpose_
medical', 'purpose_moving', 'purpose_other', 'purpose_renewable_
energy' \
        , 'purpose_small_business', 'purpose_vacation', 'purpose
_wedding']], y = y_train)
print(fitted_model.coef_)
print(fitted_model.intercept_)
```

```
[ 1.05356560e-01  4.22660385e-02  1.07861939e-01  5.
47869100e-02
  2.08162828e-01 -3.80891387e-02  8.96009682e-02 -5.
62634011e+00
 -5.48685995e+00 -3.63101119e+00 -1.83331856e+00 -5.
75191546e-01
 -1.47111104e-01 -8.26574455e-04 -4.47238389e-02 -6.
65811135e-02
  4.24941998e-02 -1.64679719e-02  1.56805060e-02  8.
18803627e-03
  1.34096469e-03  5.27648422e-03  2.17993093e-02  2.
10691161e-02
  6.77627861e-02  6.12313304e-03  1.97632330e-02  2.
04875301e-02
  1.60364781e-02]
14.24694547123541
```

In [27]:

```
Score_R2_train = fitted_model.score(X = x_train_scaled[['loan_am
nt'  , 'dti'  , 'inq_last_6mths' , 'pub_rec' , 'bc_util' , 'num_
bc_tl' , 'num_tl_op_past_12m' , 'grade_A', 'grade_B', 'grade_C',
'grade_D', 'grade_E' , 'home_ownership_MORTGAGE' \
                , 'home_ownership_OTHER' , 'home_ownership_OWN'
, \
       'home_ownership_RENT' , 'loan_status_Charged Off', 'purpo
se_credit_card', 'purpose_debt_consolidation' , 'purpose_home_im
provement', 'purpose_house', 'purpose_major_purchase', 'purpose_
medical', 'purpose_moving', 'purpose_other', 'purpose_renewable_
energy' \
        , 'purpose_small_business', 'purpose_vacation', 'purpose
_wedding']], y = y_train)
print('Training Accuracy:  ', Score_R2_train)
Score_R2 = fitted_model.score(X = x_test_scaled[['loan_amnt'  ,
'dti'  , 'inq_last_6mths' , 'pub_rec' , 'bc_util' , 'num_bc_tl'
, 'num_tl_op_past_12m' , 'grade_A', 'grade_B', 'grade_C', 'grade
_D', 'grade_E' , 'home_ownership_MORTGAGE' \
                , 'home_ownership_OTHER' , 'home_ownership_OWN'
, \
       'home_ownership_RENT' , 'loan_status_Charged Off', 'purpo
se_credit_card', 'purpose_debt_consolidation' , 'purpose_home_im
provement', 'purpose_house', 'purpose_major_purchase', 'purpose_
medical', 'purpose_moving', 'purpose_other', 'purpose_renewable_
energy' \
        , 'purpose_small_business', 'purpose_vacation', 'purpose
_wedding']], y = y_test)
print('Testing Accuracy:  ',Score_R2)
```

```
Training Accuracy:   0.9427503365430738
Testing Accuracy:   0.9432286347084489
```

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]: