0_data_preprocessing_LendingClub

0.0.1 lending club data preprocessing

Purpose Preparing lending club data for binary class classification #### overview 1. only keep rows with target classes ('Charged Off' and 'Fully Paid') 2. remove 2.1 columns with unique value 2.2 columns contain more than 40% missing values 2.3 columns which may cause informatoin leakage 2.4 columns provide redundant information 2.5 rows with too many missing values 3. Categorical variable 3.1 remove variables with too many levels 3.2 dummy coding 4. fill missing values using Multivariate imputer 5. data normalization 6. train test spliting

[1]:
```
#ignore warnings
import warnings
warnings.filterwarnings("ignore")

import pandas as pd
pd.set_option('max_columns', 100)
pd.set_option('max_colwidth', 5000) #show more information of columns

import numpy as np
from sklearn import preprocessing #data normalization
from sklearn.experimental import enable_iterative_imputer #missing value␣
↪imputation
from sklearn.impute import IterativeImputer
from sklearn.model_selection import train_test_split
```

0.0.2 data info

Since loans in Lending Club are either 36 or 60 months, and in order to have a good number of finished loans, loan data issued from 2011 to 2014 were downloaded. And in the datasets after 2015, many of the loan status are 'current' and not finished.

[2]:
```
df = pd.read_csv("LendingClub_LoanStats_2014.csv")
df.sample(5)
```

[2]:

	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	\
73368	27552291	NaN	14500.0	14500.0	14500.0	
139787	18506237	NaN	6850.0	6850.0	6850.0	
97242	23914923	NaN	4900.0	4900.0	4900.0	
234423	10069958	NaN	13000.0	13000.0	13000.0	

```
187106  12495281        NaN    12050.0      12050.0          12050.0

             term int_rate  installment grade sub_grade  \
73368    36 months   12.99%       488.50     C        C1
139787   36 months   10.15%       221.52     B        B2
97242    36 months   14.99%       169.84     C        C5
234423   60 months   16.99%       323.02     D        D1
187106   36 months   15.31%       419.55     C        C4

                 emp_title emp_length home_ownership  annual_inc  \
73368   Research Coordinator    7 years           RENT     45000.0
139787            Controller    3 years       MORTGAGE    110000.0
97242         Administration  10+ years           RENT     70000.0
234423                 E-5    5 years           RENT     50000.0
187106    Office Assistant 3     1 year           RENT     28440.0

       verification_status issue_d  loan_status pymnt_plan  \
73368         Not Verified  Sep-14  Charged Off          n
139787     Source Verified  Jun-14   Fully Paid          n
97242      Source Verified  Aug-14   Fully Paid          n
234423            Verified  Jan-14   Fully Paid          n
187106            Verified  Apr-14   Fully Paid          n

                                                                 url  \
73368   https://lendingclub.com/browse/loanDetail.action?loan_id=27552291
139787  https://lendingclub.com/browse/loanDetail.action?loan_id=18506237
97242   https://lendingclub.com/browse/loanDetail.action?loan_id=23914923
234423  https://lendingclub.com/browse/loanDetail.action?loan_id=10069958
187106  https://lendingclub.com/browse/loanDetail.action?loan_id=12495281

       desc             purpose               title zip_code addr_state  \
73368   NaN  debt_consolidation  Debt consolidation    850xx         AZ
139787  NaN  debt_consolidation  Debt consolidation    352xx         AL
97242   NaN  debt_consolidation  Debt consolidation    100xx         NY
234423  NaN         credit_card    Credit Refinance    921xx         CA
187106  NaN  debt_consolidation  Debt consolidation    985xx         WA

          dti  delinq_2yrs earliest_cr_line  fico_range_low  fico_range_high  \
73368   33.52          2.0           Jun-98           690.0            694.0
139787  26.34          0.0           Sep-00           685.0            689.0
97242   16.56          0.0           Mar-82           695.0            699.0
234423   6.79          0.0           Jun-05           680.0            684.0
187106   4.66          0.0           Jun-00           720.0            724.0

       inq_last_6mths  mths_since_last_delinq  mths_since_last_record  \
73368             1.0                    14.0                     NaN
139787            0.0                     NaN                     NaN
```

```
97242              0.0                    NaN                    NaN
234423             2.0                   56.0                    NaN
187106             2.0                    NaN                    NaN

        open_acc  pub_rec  revol_bal revol_util  total_acc  \
73368        9.0      0.0    13295.0     85.80%       36.0
139787      20.0      0.0    61559.0     59.20%       41.0
97242       12.0      0.0    32845.0     88.10%       14.0
234423       5.0      0.0    15049.0     43.40%       14.0
187106       6.0      0.0     3192.0     28.50%        6.0

       initial_list_status  out_prncp  out_prncp_inv  total_pymnt  \
73368                    f        0.0            0.0   6848.47000
139787                   w        0.0            0.0   6907.94000
97242                    f        0.0            0.0   6114.06476
234423                   f        0.0            0.0  18117.53379
187106                   f        0.0            0.0  15087.43644

        total_pymnt_inv  total_rec_prncp  total_rec_int  total_rec_late_fee  \
73368           6848.47          3481.63        1403.37                 0.0
139787          6907.94          6850.00          57.94                 0.0
97242           6114.06          4900.00        1214.06                 0.0
234423         18117.53         13000.00        5117.53                 0.0
187106         15087.44         12050.00        3037.44                 0.0

        recoveries  collection_recovery_fee last_pymnt_d  last_pymnt_amnt  \
73368      1963.47                353.4246       Aug-15           488.50
139787        0.00                  0.0000       Jul-14          6907.94
97242         0.00                  0.0000       Aug-17           169.66
234423        0.00                  0.0000       Dec-16          7134.85
187106        0.00                  0.0000       Jan-17          1661.84

       next_pymnt_d  ... num_rev_accts  num_rev_tl_bal_gt_0  num_sats  \
73368           NaN  ...           9.0                  6.0       9.0
139787          NaN  ...          17.0                 11.0      20.0
97242           NaN  ...          11.0                  9.0      12.0
234423          NaN  ...          11.0                  2.0       5.0
187106          NaN  ...           6.0                  5.0       6.0

        num_tl_120dpd_2m  num_tl_30dpd  num_tl_90g_dpd_24m num_tl_op_past_12m  \
73368                0.0           0.0                 0.0                1.0
139787               0.0           0.0                 0.0                4.0
97242                0.0           0.0                 0.0                1.0
234423               0.0           0.0                 0.0                1.0
187106               0.0           0.0                 0.0                2.0

        pct_tl_nvr_dlq  percent_bc_gt_75  pub_rec_bankruptcies  tax_liens  \
```

```
73368            96.8            100.0                  0.0        0.0
139787          100.0             40.0                  0.0        0.0
97242           100.0             83.3                  0.0        0.0
234423           78.6              NaN                  0.0        0.0
187106          100.0              0.0                  0.0        0.0

        tot_hi_cred_lim  total_bal_ex_mort  total_bc_limit  \
73368          166309.0           143617.0         12400.0
139787         460486.0           143591.0         92500.0
97242           95606.0            85340.0         31300.0
234423          34700.0            15049.0             0.0
187106          11200.0             3192.0          5700.0

        total_il_high_credit_limit  revol_bal_joint  sec_app_fico_range_low  \
73368                    150909.0              NaN                     NaN
139787                   136180.0              NaN                     NaN
97242                     58306.0              NaN                     NaN
234423                        0.0              NaN                     NaN
187106                        0.0              NaN                     NaN

        sec_app_fico_range_high  sec_app_earliest_cr_line  \
73368                       NaN                       NaN
139787                      NaN                       NaN
97242                       NaN                       NaN
234423                      NaN                       NaN
187106                      NaN                       NaN

        sec_app_inq_last_6mths  sec_app_mort_acc  sec_app_open_acc  \
73368                      NaN               NaN               NaN
139787                     NaN               NaN               NaN
97242                      NaN               NaN               NaN
234423                     NaN               NaN               NaN
187106                     NaN               NaN               NaN

        sec_app_revol_util  sec_app_open_act_il  sec_app_num_rev_accts  \
73368                  NaN                  NaN                    NaN
139787                 NaN                  NaN                    NaN
97242                  NaN                  NaN                    NaN
234423                 NaN                  NaN                    NaN
187106                 NaN                  NaN                    NaN

        sec_app_chargeoff_within_12_mths  sec_app_collections_12_mths_ex_med  \
73368                                NaN                                 NaN
139787                               NaN                                 NaN
97242                                NaN                                 NaN
234423                               NaN                                 NaN
187106                               NaN                                 NaN
```

```
        sec_app_mths_since_last_major_derog  hardship_flag  hardship_type  \
73368                                   NaN              N            NaN
139787                                  NaN              N            NaN
97242                                   NaN              N            NaN
234423                                  NaN              N            NaN
187106                                  NaN              N            NaN

        hardship_reason  hardship_status  deferral_term  hardship_amount  \
73368               NaN              NaN            NaN              NaN
139787              NaN              NaN            NaN              NaN
97242               NaN              NaN            NaN              NaN
234423              NaN              NaN            NaN              NaN
187106              NaN              NaN            NaN              NaN

        hardship_start_date  hardship_end_date  payment_plan_start_date  \
73368                   NaN                NaN                      NaN
139787                  NaN                NaN                      NaN
97242                   NaN                NaN                      NaN
234423                  NaN                NaN                      NaN
187106                  NaN                NaN                      NaN

        hardship_length  hardship_dpd  hardship_loan_status  \
73368               NaN           NaN                   NaN
139787              NaN           NaN                   NaN
97242               NaN           NaN                   NaN
234423              NaN           NaN                   NaN
187106              NaN           NaN                   NaN

        orig_projected_additional_accrued_interest  \
73368                                          NaN
139787                                         NaN
97242                                          NaN
234423                                         NaN
187106                                         NaN

        hardship_payoff_balance_amount  hardship_last_payment_amount  \
73368                              NaN                           NaN
139787                             NaN                           NaN
97242                              NaN                           NaN
234423                             NaN                           NaN
187106                             NaN                           NaN

        debt_settlement_flag  debt_settlement_flag_date  settlement_status  \
73368                      N                        NaN                NaN
139787                     N                        NaN                NaN
97242                      N                        NaN                NaN
```

```
234423                    N                        NaN                 NaN
187106                    N                        NaN                 NaN

        settlement_date  settlement_amount  settlement_percentage  \
73368               NaN                NaN                    NaN
139787              NaN                NaN                    NaN
97242               NaN                NaN                    NaN
234423              NaN                NaN                    NaN
187106              NaN                NaN                    NaN

        settlement_term
73368               NaN
139787              NaN
97242               NaN
234423              NaN
187106              NaN

[5 rows x 150 columns]
```

[3]:
```
df.shape
```

[3]:
```
(235631, 150)
```

There are 235631 observations and 150 features in this dataset

0.0.3 round1: target feature

There are 193878 Fully Paid (good) loans and 41748 Charged Off (bad) loans, since the purpose of this project is to classify good and bad loans, other loan status types will be removed at this step.

[4]:
```
df['loan_status'].value_counts()
```

[4]:
```
Fully Paid             193878
Charged Off             41748
Current                     1
Late (31-120 days)          1
Default                     1
Name: loan_status, dtype: int64
```

[5]:
```
df1 = df[(df['loan_status'] == "Fully Paid") | (df['loan_status'] == "Charged␣
 ↪Off")]
df1.shape
```

[5]:
```
(235626, 150)
```

0.0.4 round2: remove columns with unique value

Columns with only one value provide no help on classification.

[6]:
```
df1Columns = df1.columns
dropUniqueColumns = []
for col in df1Columns:
    if len(df1[col].unique()) == 1:
        dropUniqueColumns.append(col)

df2 = df1.drop(dropUniqueColumns, axis=1)

print("number of columns with unique values: " + str(len(dropUniqueColumns)))
dropUniqueColumns
```

```
number of columns with unique values: 37
```

[6]:
```
['member_id',
 'pymnt_plan',
 'out_prncp',
 'out_prncp_inv',
 'next_pymnt_d',
 'policy_code',
 'application_type',
 'annual_inc_joint',
 'dti_joint',
 'verification_status_joint',
 'open_acc_6m',
 'open_act_il',
 'open_il_12m',
 'open_il_24m',
 'mths_since_rcnt_il',
 'total_bal_il',
 'il_util',
 'open_rv_12m',
 'open_rv_24m',
 'max_bal_bc',
 'all_util',
 'inq_fi',
 'total_cu_tl',
 'inq_last_12m',
 'revol_bal_joint',
 'sec_app_fico_range_low',
 'sec_app_fico_range_high',
 'sec_app_earliest_cr_line',
 'sec_app_inq_last_6mths',
 'sec_app_mort_acc',
 'sec_app_open_acc',
 'sec_app_revol_util',
 'sec_app_open_act_il',
 'sec_app_num_rev_accts',
```

```
 'sec_app_chargeoff_within_12_mths',
 'sec_app_collections_12_mths_ex_med',
 'sec_app_mths_since_last_major_derog']
```

0.0.5 round3: remove features with more than 40% missing values

[7]:
```
df2Columns = df2.columns
dropNullColumns = []
for col in df2Columns:
    colNullPercent = (df2[col].isna().sum())/len(df2[col])
    if colNullPercent > 0.40:
        dropNullColumns.append(col)

df3 = df2.drop(dropNullColumns, axis=1)

print("number of columns with missing value more than 40%: " +␣
 ↪str(len(dropNullColumns)))
dropNullColumns
```

```
number of columns with missing value more than 40%: 26
```

[7]:
```
['desc',
 'mths_since_last_delinq',
 'mths_since_last_record',
 'mths_since_last_major_derog',
 'mths_since_recent_bc_dlq',
 'mths_since_recent_revol_delinq',
 'hardship_type',
 'hardship_reason',
 'hardship_status',
 'deferral_term',
 'hardship_amount',
 'hardship_start_date',
 'hardship_end_date',
 'payment_plan_start_date',
 'hardship_length',
 'hardship_dpd',
 'hardship_loan_status',
 'orig_projected_additional_accrued_interest',
 'hardship_payoff_balance_amount',
 'hardship_last_payment_amount',
 'debt_settlement_flag_date',
 'settlement_status',
 'settlement_date',
 'settlement_amount',
 'settlement_percentage',
 'settlement_term']
```

take a look at the remaining data after first 3 rounds, number of features are reduced from 150 to 87. And some of them still have missing value, will deal with them later.

[8]:
```
df3.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 235626 entries, 0 to 235628
Data columns (total 87 columns):
 #   Column                      Non-Null Count   Dtype
---  ------                      --------------   -----
 0   id                          235626 non-null  object
 1   loan_amnt                   235626 non-null  float64
 2   funded_amnt                 235626 non-null  float64
 3   funded_amnt_inv             235626 non-null  float64
 4   term                        235626 non-null  object
 5   int_rate                    235626 non-null  object
 6   installment                 235626 non-null  float64
 7   grade                       235626 non-null  object
 8   sub_grade                   235626 non-null  object
 9   emp_title                   222390 non-null  object
 10  emp_length                  223607 non-null  object
 11  home_ownership              235626 non-null  object
 12  annual_inc                  235626 non-null  float64
 13  verification_status         235626 non-null  object
 14  issue_d                     235626 non-null  object
 15  loan_status                 235626 non-null  object
 16  url                         235626 non-null  object
 17  purpose                     235626 non-null  object
 18  title                       235626 non-null  object
 19  zip_code                    235626 non-null  object
 20  addr_state                  235626 non-null  object
 21  dti                         235626 non-null  float64
 22  delinq_2yrs                 235626 non-null  float64
 23  earliest_cr_line            235626 non-null  object
 24  fico_range_low              235626 non-null  float64
 25  fico_range_high             235626 non-null  float64
 26  inq_last_6mths              235626 non-null  float64
 27  open_acc                    235626 non-null  float64
 28  pub_rec                     235626 non-null  float64
 29  revol_bal                   235626 non-null  float64
 30  revol_util                  235501 non-null  object
 31  total_acc                   235626 non-null  float64
 32  initial_list_status         235626 non-null  object
 33  total_pymnt                 235626 non-null  float64
 34  total_pymnt_inv             235626 non-null  float64
 35  total_rec_prncp             235626 non-null  float64
 36  total_rec_int               235626 non-null  float64
 37  total_rec_late_fee          235626 non-null  float64
```

```
38  recoveries                  235626 non-null  float64
39  collection_recovery_fee     235626 non-null  float64
40  last_pymnt_d                235483 non-null  object
41  last_pymnt_amnt             235626 non-null  float64
42  last_credit_pull_d          235600 non-null  object
43  last_fico_range_high        235626 non-null  float64
44  last_fico_range_low         235626 non-null  float64
45  collections_12_mths_ex_med  235626 non-null  float64
46  acc_now_delinq              235626 non-null  float64
47  tot_coll_amt                235626 non-null  float64
48  tot_cur_bal                 235626 non-null  float64
49  total_rev_hi_lim            235626 non-null  float64
50  acc_open_past_24mths        235626 non-null  float64
51  avg_cur_bal                 235620 non-null  float64
52  bc_open_to_buy              233181 non-null  float64
53  bc_util                     233015 non-null  float64
54  chargeoff_within_12_mths    235626 non-null  float64
55  delinq_amnt                 235626 non-null  float64
56  mo_sin_old_il_acct          228455 non-null  float64
57  mo_sin_old_rev_tl_op        235626 non-null  float64
58  mo_sin_rcnt_rev_tl_op       235626 non-null  float64
59  mo_sin_rcnt_tl              235626 non-null  float64
60  mort_acc                    235626 non-null  float64
61  mths_since_recent_bc        233380 non-null  float64
62  mths_since_recent_inq       213934 non-null  float64
63  num_accts_ever_120_pd       235626 non-null  float64
64  num_actv_bc_tl              235626 non-null  float64
65  num_actv_rev_tl             235626 non-null  float64
66  num_bc_sats                 235626 non-null  float64
67  num_bc_tl                   235626 non-null  float64
68  num_il_tl                   235626 non-null  float64
69  num_op_rev_tl               235626 non-null  float64
70  num_rev_accts               235626 non-null  float64
71  num_rev_tl_bal_gt_0         235626 non-null  float64
72  num_sats                    235626 non-null  float64
73  num_tl_120dpd_2m            227766 non-null  float64
74  num_tl_30dpd                235626 non-null  float64
75  num_tl_90g_dpd_24m          235626 non-null  float64
76  num_tl_op_past_12m          235626 non-null  float64
77  pct_tl_nvr_dlq              235626 non-null  float64
78  percent_bc_gt_75            233069 non-null  float64
79  pub_rec_bankruptcies        235626 non-null  float64
80  tax_liens                   235626 non-null  float64
81  tot_hi_cred_lim             235626 non-null  float64
82  total_bal_ex_mort           235626 non-null  float64
83  total_bc_limit              235626 non-null  float64
84  total_il_high_credit_limit  235626 non-null  float64
85  hardship_flag               235625 non-null  object
```

```
 86  debt_settlement_flag        235626 non-null  object
dtypes: float64(64), object(23)
memory usage: 158.2+ MB
```

0.0.6 round 4: mannually remove features based on Dictionary

It is needed to remove features which may cause information leakage of the final result. Since this project's purpose is to predict whether a loan will be charged off **before** issueing a loan, such features may include information that happended **after** a loan was issued. There are also many features provide similar information which are redundant and also need to be removed.

- Prepare a dataset which contains feature name, feature type, a value of the feature and the feature's description.

[9]:
```
dictionary = pd.read_csv("LCDataDictionary.csv",skipfooter=1) #skip last row
dictionary.sample(3)
```

[9]:
```
             LoanStatNew  \
20             emp_title
122  sec_app_revol_util
99             total_acc

Description  \
20                      The job title supplied by the Borrower when applying for
the loan.*
122   Ratio of total current balance to high credit/credit limit for all
revolving accounts
99                 The total number of credit lines currently in the borrower's
credit file

     Unnamed: 2  Unnamed: 3  Unnamed: 4  Unnamed: 5  Unnamed: 6  Unnamed: 7  \
20          NaN         NaN         NaN         NaN         NaN         NaN
122         NaN         NaN         NaN         NaN         NaN         NaN
99          NaN         NaN         NaN         NaN         NaN         NaN

     Unnamed: 8  Unnamed: 9  Unnamed: 10
20          NaN         NaN          NaN
122         NaN         NaN          NaN
99          NaN         NaN          NaN
```

[10]:
```
dictionary = dictionary.rename(columns={'LoanStatNew': 'name'})
dictionary = dictionary[['name','Description']]

df3Types = pd.DataFrame(df3.dtypes,columns=['dtypes']).reset_index()
df3Types['name'] = df3Types['index']
df3Types = df3Types[['name','dtypes']]
df3Types['first value'] = df3.loc[0].values
```

```
preview = df3Types.merge(dictionary, on='name',how='left')
preview.tail()
```

[10]:
```
                          name   dtypes first value  \
82           total_bal_ex_mort  float64       15030
83              total_bc_limit  float64       13000
84  total_il_high_credit_limit  float64       11325
85               hardship_flag   object           N
86        debt_settlement_flag   object           N

           Description
82                                                    Total credit
balance excluding mortgage
83                                                    Total bankcard
high credit/credit limit
84                                                 Total installment
high credit/credit limit
85                                     Flags whether or not the borrower
is on a hardship plan
86  Flags whether or not the borrower, who has charged-off, is working with a
debt-settlement company.
```

- Since there are too much features and it is hard to figure out features at once, they are divided into four groups.

round4.1 group1

[11]:
```
preview[:22]
```

[11]:
```
                   name   dtypes  \
0                    id   object
1            loan_amnt  float64
2          funded_amnt  float64
3      funded_amnt_inv  float64
4                 term   object
5             int_rate   object
6          installment  float64
7                grade   object
8            sub_grade   object
9            emp_title   object
10          emp_length   object
11      home_ownership   object
12          annual_inc  float64
13  verification_status   object
14              issue_d   object
15         loan_status   object
16                 url   object
17             purpose   object
```

```
18                title   object
19             zip_code   object
20           addr_state   object
21                  dti  float64

                                                        first value  \
0                                                          36805548
1                                                             10400
2                                                             10400
3                                                             10400
4                                                          36 months
5                                                              6.99%
6                                                             321.08
7                                                                  A
8                                                                 A3
9                                       Truck Driver Delivery Personel
10                                                          8 years
11                                                         MORTGAGE
12                                                            58000
13                                                    Not Verified
14                                                          Dec-14
15                                                     Charged Off
16  https://lendingclub.com/browse/loanDetail.action?loan_id=36805548
17                                                     credit_card
18                                          Credit card refinancing
19                                                            937xx
20                                                               CA
21                                                            14.92

                                Description
0
A unique LC assigned ID for the loan listing.
1                               The listed amount of the loan applied for by the
borrower. If at some point in time, the credit department reduces the loan
amount, then it will be reflected in this value.
2
The total amount committed to that loan at that point in time.
3
The total amount committed by investors for that loan at that point in time.
4
The number of payments on the loan. Values are in months and can be either 36 or
60.
5
Interest Rate on the loan
6
The monthly payment owed by the borrower if the loan originates.
7
```

```
LC assigned loan grade
8
LC assigned loan subgrade
9
The job title supplied by the Borrower when applying for the loan.*
10
Employment length in years. Possible values are between 0 and 10 where 0 means
less than one year and 10 means ten or more years.
11                                                    The home ownership status
provided by the borrower during registrationor obtained from the credit
report.Our values are: RENT, OWN, MORTGAGE, OTHER
12
The self-reported annual income provided by the borrower during registration.
13
Indicates if income was verified by LC, not verified, or if the income source
was verified
14
The month which the loan was funded
15
Current status of the loan
16
URL for the LC page with listing data.
17
A category provided by the borrower for the loan request.
18
The loan title provided by the borrower
19
The first 3 numbers of the zip code provided by the borrower in the loan
application.
20
The state provided by the borrower in the loan application
21  A ratio calculated using the borrowers total monthly debt payments on the
total debt obligations, excluding mortgage and the requested LC loan, divided by
the borrowers self-reported monthly income.
```

- **id, url**: useless for modeling
- **funded_amnt, funded_amnt_inv, issue_d**: happend after loan issuing
- **grade, sub_grade** redundant information as they are based on interest rate
- **emp_title, zip_code, addr_state**: too much levels based on experience

[12]:
```
drop_list = ['id','funded_amnt','funded_amnt_inv','grade','sub_grade',
             'emp_title','issue_d','url','zip_code','addr_state']
df4_1 = df3.drop(drop_list,axis=1)
df4_1.head()
```

[12]:

	loan_amnt	term	int_rate	installment	emp_length	home_ownership	\
0	10400.0	36 months	6.99%	321.08	8 years	MORTGAGE	

```
1    15000.0   60 months   12.39%       336.64  10+ years           RENT
2     9600.0   36 months   13.66%       326.53  10+ years           RENT
3    12800.0   60 months   17.14%       319.08  10+ years       MORTGAGE
4    21425.0   60 months   15.59%       516.36    6 years           RENT

   annual_inc verification_status  loan_status             purpose  \
0     58000.0        Not Verified  Charged Off         credit_card
1     78000.0     Source Verified   Fully Paid  debt_consolidation
2     69000.0     Source Verified   Fully Paid  debt_consolidation
3    125000.0            Verified   Fully Paid                 car
4     63800.0     Source Verified   Fully Paid         credit_card

                    title    dti  delinq_2yrs earliest_cr_line  \
0  Credit card refinancing  14.92          0.0           Sep-89
1       Debt consolidation  12.03          0.0           Aug-94
2       Debt consolidation  25.81          0.0           Nov-92
3            Car financing   8.31          1.0           Oct-00
4  Credit card refinancing  18.49          0.0           Aug-03

   fico_range_low  fico_range_high  inq_last_6mths  open_acc  pub_rec  \
0           710.0            714.0             2.0      17.0      0.0
1           750.0            754.0             0.0       6.0      0.0
2           680.0            684.0             0.0      12.0      0.0
3           665.0            669.0             0.0       8.0      0.0
4           685.0            689.0             0.0      10.0      0.0

   revol_bal revol_util  total_acc initial_list_status  total_pymnt  \
0     6133.0     31.60%       36.0                   w   6611.69000
1   138008.0        29%       17.0                   w  17392.37000
2    16388.0     59.40%       44.0                   f   9973.43000
3     5753.0    100.90%       13.0                   w  19165.35192
4    16374.0     76.20%       35.0                   w  25512.20000

   total_pymnt_inv  total_rec_prncp  total_rec_int  total_rec_late_fee  \
0          6611.69          5217.75         872.67                 0.0
1         17392.37         15000.00        2392.37                 0.0
2          9973.43          9600.00         373.43                 0.0
3         19165.35         12800.00        6365.35                 0.0
4         25512.20         21425.00        4087.20                 0.0

   recoveries  collection_recovery_fee last_pymnt_d  last_pymnt_amnt  \
0      521.27                  93.8286       Aug-16           321.08
1        0.00                   0.0000       Jun-16         12017.81
2        0.00                   0.0000       Apr-15          9338.58
3        0.00                   0.0000       Sep-19          1576.08
4        0.00                   0.0000       May-16         17813.19
```

```
  last_credit_pull_d  last_fico_range_high  last_fico_range_low  \
0             Feb-17                564.0                560.0
1             Jul-20                714.0                710.0
2             Jul-20                714.0                710.0
3             Apr-20                704.0                700.0
4             Apr-18                529.0                525.0

   collections_12_mths_ex_med  acc_now_delinq  tot_coll_amt  tot_cur_bal  \
0                         0.0             0.0           0.0     162110.0
1                         0.0             0.0           0.0     149140.0
2                         0.0             0.0           0.0      38566.0
3                         0.0             0.0           0.0     261815.0
4                         0.0             0.0           0.0      42315.0

   total_rev_hi_lim  acc_open_past_24mths  avg_cur_bal  bc_open_to_buy  \
0           19400.0                   7.0       9536.0          7599.0
1          184500.0                   5.0      29828.0          9525.0
2           27600.0                   8.0       3214.0          6494.0
3            5700.0                   2.0      32727.0             0.0
4           21500.0                   4.0       4232.0           324.0

   bc_util  chargeoff_within_12_mths  delinq_amnt  mo_sin_old_il_acct  \
0     41.5                       0.0          0.0                76.0
1      4.7                       0.0          0.0               103.0
2     69.2                       0.0          0.0               183.0
3    103.2                       0.0          0.0                16.0
4     97.8                       0.0          0.0               135.0

   mo_sin_old_rev_tl_op  mo_sin_rcnt_rev_tl_op  mo_sin_rcnt_tl  mort_acc  \
0                 290.0                    1.0             1.0       1.0
1                 244.0                    1.0             1.0       0.0
2                 265.0                   23.0             3.0       0.0
3                 170.0                   21.0            16.0       5.0
4                 136.0                    7.0             7.0       0.0

   mths_since_recent_bc  mths_since_recent_inq  num_accts_ever_120_pd  \
0                   5.0                    1.0                    4.0
1                  47.0                    NaN                    0.0
2                  24.0                   17.0                    0.0
3                  21.0                    1.0                    1.0
4                   7.0                    7.0                    1.0

   num_actv_bc_tl  num_actv_rev_tl  num_bc_sats  num_bc_tl  num_il_tl  \
0             6.0              9.0          7.0       18.0        2.0
1             1.0              4.0          1.0        2.0        8.0
2             4.0              7.0          5.0       16.0       17.0
3             3.0              5.0          3.0        5.0        1.0
```

```
4             3.0              4.0          3.0       12.0       16.0

   num_op_rev_tl  num_rev_accts  num_rev_tl_bal_gt_0  num_sats  \
0           14.0           32.0                  9.0      17.0
1            5.0            9.0                  4.0       6.0
2            8.0           26.0                  7.0      12.0
3            5.0            7.0                  5.0       8.0
4            5.0           18.0                  4.0      10.0

   num_tl_120dpd_2m  num_tl_30dpd  num_tl_90g_dpd_24m  num_tl_op_past_12m  \
0               0.0           0.0                 0.0                 4.0
1               0.0           0.0                 0.0                 4.0
2               0.0           0.0                 0.0                 3.0
3               0.0           0.0                 0.0                 0.0
4               0.0           0.0                 0.0                 2.0

   pct_tl_nvr_dlq  percent_bc_gt_75  pub_rec_bankruptcies  tax_liens  \
0            83.3              14.3                   0.0        0.0
1           100.0               0.0                   0.0        0.0
2           100.0              60.0                   0.0        0.0
3            76.9             100.0                   0.0        0.0
4            91.4             100.0                   0.0        0.0

   tot_hi_cred_lim  total_bal_ex_mort  total_bc_limit  \
0         179407.0            15030.0         13000.0
1         196500.0           149140.0         10000.0
2          52490.0            38566.0         21100.0
3         368700.0            18007.0          4400.0
4          57073.0            42315.0         15000.0

   total_il_high_credit_limit hardship_flag debt_settlement_flag
0                     11325.0             N                    N
1                     12000.0             N                    N
2                     24890.0             N                    N
3                     18000.0             N                    N
4                     35573.0             N                    N
```

round4.2 group2

[13]:
```
preview[22:44]
```

[13]:
```
                    name   dtypes first value  \
22             delinq_2yrs  float64           0
23         earliest_cr_line   object      Sep-89
24           fico_range_low  float64         710
25          fico_range_high  float64         714
26           inq_last_6mths  float64           2
```

27	open_acc	float64	17	
28	pub_rec	float64	0	
29	revol_bal	float64	6133	
30	revol_util	object	31.60%	
31	total_acc	float64	36	
32	initial_list_status	object	w	
33	total_pymnt	float64	6611.69	
34	total_pymnt_inv	float64	6611.69	
35	total_rec_prncp	float64	5217.75	
36	total_rec_int	float64	872.67	
37	total_rec_late_fee	float64	0	
38	recoveries	float64	521.27	
39	collection_recovery_fee	float64	93.8286	
40	last_pymnt_d	object	Aug-16	
41	last_pymnt_amnt	float64	321.08	
42	last_credit_pull_d	object	Feb-17	
43	last_fico_range_high	float64	564	

	Description
22	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
23	The month the borrower's earliest reported credit line was opened
24	The lower boundary range the borrowers FICO at loan origination belongs to.
25	The upper boundary range the borrowers FICO at loan origination belongs to.
26	The number of inquiries in past 6 months (excluding auto and mortgage inquiries)
27	The number of open credit lines in the borrower's credit file.
28	Number of derogatory public records
29	Total credit revolving balance
30	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
31	The total number of credit lines currently in the borrower's credit file
32	The initial listing status of the loan. Possible values are W, F
33	Payments received to date for total amount funded
34	Payments received to date for portion of total amount funded by investors
35	Principal received to date

```
36
Interest received to date
37
Late fees received to date
38
post charge off gross recovery
39
post charge off collection fee
40
Last month payment was received
41
Last total payment amount received
42                                                                    The
most recent month LC pulled credit for this loan
43                                                    The upper boundary
range the borrowers last FICO pulled belongs to.
```

- **total_pymnt, total_pymnt_inv, total_rec_prncp, total_rec_int, total_rec_late_fee, recoveries, collection_recovery_fee, last_pymnt_d, last_pymnt_amnt**: happend after loan issuing

[14]:
```
drop_list = ['total_pymnt','total_pymnt_inv',
             'total_rec_prncp',
             'total_rec_int','total_rec_late_fee',
             'recoveries','collection_recovery_fee',
             'last_pymnt_d','last_pymnt_amnt']
df4_2 = df4_1.drop(drop_list,axis=1)
df4_2.head()
```

[14]:

	loan_amnt	term	int_rate	installment	emp_length	home_ownership	annual_inc	verification_status	loan_status	purpose	title	dti	delinq_2yrs	earliest_cr_line
0	10400.0	36 months	6.99%	321.08	8 years	MORTGAGE	58000.0	Not Verified	Charged Off	credit_card	Credit card refinancing	14.92	0.0	Sep-89
1	15000.0	60 months	12.39%	336.64	10+ years	RENT	78000.0	Source Verified	Fully Paid	debt_consolidation	Debt consolidation	12.03	0.0	Aug-94
2	9600.0	36 months	13.66%	326.53	10+ years	RENT	69000.0	Source Verified	Fully Paid	debt_consolidation	Debt consolidation	25.81	0.0	Nov-92
3	12800.0	60 months	17.14%	319.08	10+ years	MORTGAGE	125000.0	Verified	Fully Paid	car	Car financing	8.31	1.0	Oct-00
4	21425.0	60 months	15.59%	516.36	6 years	RENT	63800.0	Source Verified	Fully Paid	credit_card				

```
4  Credit card refinancing  18.49          0.0           Aug-03

   fico_range_low  fico_range_high  inq_last_6mths  open_acc  pub_rec  \
0           710.0            714.0             2.0      17.0      0.0
1           750.0            754.0             0.0       6.0      0.0
2           680.0            684.0             0.0      12.0      0.0
3           665.0            669.0             0.0       8.0      0.0
4           685.0            689.0             0.0      10.0      0.0

   revol_bal revol_util  total_acc initial_list_status last_credit_pull_d  \
0     6133.0     31.60%       36.0                   w             Feb-17
1   138008.0        29%       17.0                   w             Jul-20
2    16388.0     59.40%       44.0                   f             Jul-20
3     5753.0    100.90%       13.0                   w             Apr-20
4    16374.0     76.20%       35.0                   w             Apr-18

   last_fico_range_high  last_fico_range_low  collections_12_mths_ex_med  \
0                 564.0                560.0                         0.0
1                 714.0                710.0                         0.0
2                 714.0                710.0                         0.0
3                 704.0                700.0                         0.0
4                 529.0                525.0                         0.0

   acc_now_delinq  tot_coll_amt  tot_cur_bal  total_rev_hi_lim  \
0             0.0           0.0     162110.0           19400.0
1             0.0           0.0     149140.0          184500.0
2             0.0           0.0      38566.0           27600.0
3             0.0           0.0     261815.0            5700.0
4             0.0           0.0      42315.0           21500.0

   acc_open_past_24mths  avg_cur_bal  bc_open_to_buy  bc_util  \
0                   7.0       9536.0          7599.0     41.5
1                   5.0      29828.0          9525.0      4.7
2                   8.0       3214.0          6494.0     69.2
3                   2.0      32727.0             0.0    103.2
4                   4.0       4232.0           324.0     97.8

   chargeoff_within_12_mths  delinq_amnt  mo_sin_old_il_acct  \
0                       0.0          0.0                76.0
1                       0.0          0.0               103.0
2                       0.0          0.0               183.0
3                       0.0          0.0                16.0
4                       0.0          0.0               135.0

   mo_sin_old_rev_tl_op  mo_sin_rcnt_rev_tl_op  mo_sin_rcnt_tl  mort_acc  \
0                 290.0                    1.0             1.0       1.0
1                 244.0                    1.0             1.0       0.0
```

```
2                265.0                  23.0            3.0       0.0
3                170.0                  21.0           16.0       5.0
4                136.0                   7.0            7.0       0.0

   mths_since_recent_bc  mths_since_recent_inq  num_accts_ever_120_pd  \
0                   5.0                    1.0                    4.0
1                  47.0                    NaN                    0.0
2                  24.0                   17.0                    0.0
3                  21.0                    1.0                    1.0
4                   7.0                    7.0                    1.0

   num_actv_bc_tl  num_actv_rev_tl  num_bc_sats  num_bc_tl  num_il_tl  \
0             6.0              9.0          7.0       18.0        2.0
1             1.0              4.0          1.0        2.0        8.0
2             4.0              7.0          5.0       16.0       17.0
3             3.0              5.0          3.0        5.0        1.0
4             3.0              4.0          3.0       12.0       16.0

   num_op_rev_tl  num_rev_accts  num_rev_tl_bal_gt_0  num_sats  \
0           14.0           32.0                  9.0      17.0
1            5.0            9.0                  4.0       6.0
2            8.0           26.0                  7.0      12.0
3            5.0            7.0                  5.0       8.0
4            5.0           18.0                  4.0      10.0

   num_tl_120dpd_2m  num_tl_30dpd  num_tl_90g_dpd_24m  num_tl_op_past_12m  \
0               0.0           0.0                 0.0                 4.0
1               0.0           0.0                 0.0                 4.0
2               0.0           0.0                 0.0                 3.0
3               0.0           0.0                 0.0                 0.0
4               0.0           0.0                 0.0                 2.0

   pct_tl_nvr_dlq  percent_bc_gt_75  pub_rec_bankruptcies  tax_liens  \
0            83.3              14.3                   0.0        0.0
1           100.0               0.0                   0.0        0.0
2           100.0              60.0                   0.0        0.0
3            76.9             100.0                   0.0        0.0
4            91.4             100.0                   0.0        0.0

   tot_hi_cred_lim  total_bal_ex_mort  total_bc_limit  \
0         179407.0            15030.0         13000.0
1         196500.0           149140.0         10000.0
2          52490.0            38566.0         21100.0
3         368700.0            18007.0          4400.0
4          57073.0            42315.0         15000.0

   total_il_high_credit_limit hardship_flag debt_settlement_flag
```

```
0                    11325.0            N                    N
1                    12000.0            N                    N
2                    24890.0            N                    N
3                    18000.0            N                    N
4                    35573.0            N                    N
```

round4.3 group3

[15]:
```
preview[44:66]
```

[15]:
```
                          name   dtypes first value  \
44         last_fico_range_low  float64         560
45  collections_12_mths_ex_med  float64           0
46              acc_now_delinq  float64           0
47                tot_coll_amt  float64           0
48                 tot_cur_bal  float64      162110
49            total_rev_hi_lim  float64       19400
50        acc_open_past_24mths  float64           7
51                 avg_cur_bal  float64        9536
52              bc_open_to_buy  float64        7599
53                     bc_util  float64        41.5
54    chargeoff_within_12_mths  float64           0
55                 delinq_amnt  float64           0
56          mo_sin_old_il_acct  float64          76
57        mo_sin_old_rev_tl_op  float64         290
58       mo_sin_rcnt_rev_tl_op  float64           1
59              mo_sin_rcnt_tl  float64           1
60                    mort_acc  float64           1
61        mths_since_recent_bc  float64           5
62       mths_since_recent_inq  float64           1
63       num_accts_ever_120_pd  float64           4
64              num_actv_bc_tl  float64           6
65             num_actv_rev_tl  float64           9

Description
44                   The lower boundary range the borrowers last FICO pulled
belongs to.
45                       Number of collections in 12 months excluding medical
collections
46                        The number of accounts on which the borrower is now
delinquent.
47                                                 Total collection amounts
ever owed
48                                               Total current balance of all
accounts
49
NaN
```

```
50                                                 Number of trades opened in past
24 months.
51                                                    Average current balance of all
accounts
52                                                  Total open to buy on revolving
bankcards.
53  Ratio of total current balance to high credit/credit limit for all bankcard
accounts.
54                                                   Number of charge-offs within
12 months
55     The past-due amount owed for the accounts on which the borrower is now
delinquent.
56                                       Months since oldest bank installment
account opened
57                                             Months since oldest revolving
account opened
58                                        Months since most recent revolving
account opened
59                                                Months since most recent
account opened
60                                                                Number of mortgage
accounts.
61                                        Months since most recent bankcard
account opened.
62                                                            Months since most recent
inquiry.
63                                        Number of accounts ever 120 or more days
past due
64                                             Number of currently active bankcard
accounts
65                                              Number of currently active
revolving trades
```

round4.4 group4

[16]:
```
preview[66:]
```

[16]:

	name	dtypes	first value	\
66	num_bc_sats	float64	7	
67	num_bc_tl	float64	18	
68	num_il_tl	float64	2	
69	num_op_rev_tl	float64	14	
70	num_rev_accts	float64	32	
71	num_rev_tl_bal_gt_0	float64	9	
72	num_sats	float64	17	
73	num_tl_120dpd_2m	float64	0	
74	num_tl_30dpd	float64	0	

```
75          num_tl_90g_dpd_24m  float64           0
76          num_tl_op_past_12m  float64           4
77              pct_tl_nvr_dlq  float64        83.3
78            percent_bc_gt_75  float64        14.3
79        pub_rec_bankruptcies  float64           0
80                   tax_liens  float64           0
81             tot_hi_cred_lim  float64      179407
82           total_bal_ex_mort  float64       15030
83              total_bc_limit  float64       13000
84  total_il_high_credit_limit  float64       11325
85               hardship_flag   object           N
86        debt_settlement_flag   object           N

           Description
66                                                    Number of
satisfactory bankcard accounts
67
Number of bankcard accounts
68                                                       Number
of installment accounts
69                                                    Number of
open revolving accounts
70                                                       Number
of revolving accounts
71                                                 Number of revolving
trades with balance >0
72                                                    Number of
satisfactory accounts
73                           Number of accounts currently 120 days past due
(updated in past 2 months)
74                            Number of accounts currently 30 days past due
(updated in past 2 months)
75                                   Number of accounts 90 or more days past
due in last 24 months
76                                               Number of accounts
opened in past 12 months
77                                                    Percent of
trades never delinquent
78                                          Percentage of all bankcard
accounts > 75% of limit.
79                                                    Number of
public record bankruptcies
80
Number of tax liens
81                                                       Total
high credit/credit limit
82                                                  Total credit
```

```
balance excluding mortgage
83                                                    Total bankcard
high credit/credit limit
84                                                 Total installment
high credit/credit limit
85                                          Flags whether or not the borrower
is on a hardship plan
86  Flags whether or not the borrower, who has charged-off, is working with a
debt-settlement company.
```

- no features need to remove after round 4.3 and 4.4

0.0.7 round5 missing value

- There are 68 features remaining

[17]:
```
df4_2.shape
```

[17]:
```
(235626, 68)
```

- Keep only the rows with at least 66 non-NA values (remove rows with 2 or more missing values)

[18]:
```
df5 = df4_2.dropna(thresh=66)
df5.shape
```

[18]:
```
(232885, 68)
```

- check remaining missing value

[19]:
```
null_counts = df5.isnull().sum()
print(null_counts[null_counts>0])
```

```
emp_length              11571
revol_util                 23
last_credit_pull_d         24
bc_util                   159
mo_sin_old_il_acct       6887
mths_since_recent_inq   21177
num_tl_120dpd_2m         7738
percent_bc_gt_75           83
hardship_flag               1
dtype: int64
```

- revol_util, last_credit_pull_d, etc only have few missing values, remove rows have missing values in them

[20]:
```
df5 = df5.
 ↪dropna(subset=['revol_util','last_credit_pull_d','bc_util','percent_bc_gt_75','hardship_flag'
```

```
null_counts = df5.isnull().sum()
print(null_counts[null_counts>0])
```

```
emp_length               11561
mo_sin_old_il_acct        6884
mths_since_recent_inq    21167
num_tl_120dpd_2m          7647
dtype: int64
```

0.0.8 round 6.1 Categorical variables

'object' type may contain categorical variables.

[21]:
```
object_columns_df = df5.select_dtypes(include=['object'])
print(object_columns_df.iloc[0])
```

```
term                                 36 months
int_rate                                 6.99%
emp_length                             8 years
home_ownership                        MORTGAGE
verification_status               Not Verified
loan_status                        Charged Off
purpose                            credit_card
title                  Credit card refinancing
earliest_cr_line                        Sep-89
revol_util                              31.60%
initial_list_status                          w
last_credit_pull_d                      Feb-17
hardship_flag                                N
debt_settlement_flag                         N
Name: 0, dtype: object
```

- remove **earliest_cr_line** and **last_credit_pull_d** because they are date info and are useless for modeling
- correct **int_rate** and **revol_util** to float type

[22]:
```
df6_1 = df5.drop(['earliest_cr_line','last_credit_pull_d'], axis=1)
for col in ['int_rate','revol_util']:
    df6_1[col] = df6_1[col].str.rstrip('%').astype('float')
```

- check distribution of categorical variables

[23]:
```
cols = ['loan_status','term','emp_length','home_ownership','verification_status',
       ␣
 ↪'purpose','title','initial_list_status','hardship_flag','debt_settlement_flag']
for col in cols:
    print(df6_1[col].value_counts(dropna=False))
    print("-------")
```

```
Fully Paid     191439
Charged Off     41179
Name: loan_status, dtype: int64
-------
 36 months    160328
 60 months     72290
Name: term, dtype: int64
-------
10+ years    78446
2 years      20282
3 years      18087
< 1 year     17804
1 year       14449
4 years      13384
7 years      12951
5 years      12920
8 years      11726
6 years      11675
NaN          11561
9 years       9333
Name: emp_length, dtype: int64
-------
MORTGAGE    118392
RENT         91573
OWN          22652
ANY              1
Name: home_ownership, dtype: int64
-------
Source Verified    98736
Not Verified       67622
Verified           66260
Name: verification_status, dtype: int64
-------
debt_consolidation    141268
credit_card            55111
home_improvement       12738
other                  10106
major_purchase          3781
medical                 2265
small_business          2236
car                     1799
moving                  1294
vacation                1153
house                    738
renewable_energy         121
wedding                    8
Name: purpose, dtype: int64
-------
```

```
Debt consolidation                          138908
Credit card refinancing                      53944
Home improvement                             12578
Other                                         9972
Major purchase                                3740
                                             ...
personal freedom                                 1
credit pay off                                   1
Careful Preparation Leads to Success             1
to pay off at a lower rate that am now           1
NEW YEARS LIFE                                   1
Name: title, Length: 2040, dtype: int64
-------
w    121952
f    110666
Name: initial_list_status, dtype: int64
-------
N    232618
Name: hardship_flag, dtype: int64
-------
N    228109
Y      4509
Name: debt_settlement_flag, dtype: int64
-------
```

remove

- **title** has too much levels
- **title** only have one level
- **home_ownership** with level 'ANY'

[24]:
```
df6_1 = df6_1.drop(['title','hardship_flag'], axis=1)
df6_1 = df6_1.drop(df6_1[df6_1['home_ownership'] == 'ANY'].index)
```

map

- **emp_length** to ordinal numbers
- **loan_status** to 0, 1

[25]:
```
mapping_dict = {
    'emp_length':{
        '10+ years': 10,
        '9 years': 9,
        '8 years': 8,
        '7 years': 7,
        '6 years': 6,
        '5 years': 5,
        '4 years': 4,
```

```
        '3 years': 3,
        '2 years': 2,
        '1 year': 1,
        '< 1 year': 0
    },
    'loan_status':{
        'Fully Paid':0,
        'Charged Off':1
    }
}

df6_1 = df6_1.replace(mapping_dict)
df6_1.head()
```

[25]:

	loan_amnt	term	int_rate	installment	emp_length	home_ownership
0	10400.0	36 months	6.99	321.08	8.0	MORTGAGE
1	15000.0	60 months	12.39	336.64	10.0	RENT
2	9600.0	36 months	13.66	326.53	10.0	RENT
3	12800.0	60 months	17.14	319.08	10.0	MORTGAGE
4	21425.0	60 months	15.59	516.36	6.0	RENT

	annual_inc	verification_status	loan_status	purpose	dti
0	58000.0	Not Verified	1	credit_card	14.92
1	78000.0	Source Verified	0	debt_consolidation	12.03
2	69000.0	Source Verified	0	debt_consolidation	25.81
3	125000.0	Verified	0	car	8.31
4	63800.0	Source Verified	0	credit_card	18.49

	delinq_2yrs	fico_range_low	fico_range_high	inq_last_6mths	open_acc
0	0.0	710.0	714.0	2.0	17.0
1	0.0	750.0	754.0	0.0	6.0
2	0.0	680.0	684.0	0.0	12.0
3	1.0	665.0	669.0	0.0	8.0
4	0.0	685.0	689.0	0.0	10.0

	pub_rec	revol_bal	revol_util	total_acc	initial_list_status
0	0.0	6133.0	31.6	36.0	w
1	0.0	138008.0	29.0	17.0	w
2	0.0	16388.0	59.4	44.0	f
3	0.0	5753.0	100.9	13.0	w
4	0.0	16374.0	76.2	35.0	w

	last_fico_range_high	last_fico_range_low	collections_12_mths_ex_med
0	564.0	560.0	0.0
1	714.0	710.0	0.0
2	714.0	710.0	0.0
3	704.0	700.0	0.0

```
4                529.0               525.0                         0.0

   acc_now_delinq  tot_coll_amt  tot_cur_bal  total_rev_hi_lim  \
0             0.0           0.0     162110.0           19400.0
1             0.0           0.0     149140.0          184500.0
2             0.0           0.0      38566.0           27600.0
3             0.0           0.0     261815.0            5700.0
4             0.0           0.0      42315.0           21500.0

   acc_open_past_24mths  avg_cur_bal  bc_open_to_buy  bc_util  \
0                   7.0       9536.0          7599.0     41.5
1                   5.0      29828.0          9525.0      4.7
2                   8.0       3214.0          6494.0     69.2
3                   2.0      32727.0             0.0    103.2
4                   4.0       4232.0           324.0     97.8

   chargeoff_within_12_mths  delinq_amnt  mo_sin_old_il_acct  \
0                       0.0          0.0                76.0
1                       0.0          0.0               103.0
2                       0.0          0.0               183.0
3                       0.0          0.0                16.0
4                       0.0          0.0               135.0

   mo_sin_old_rev_tl_op  mo_sin_rcnt_rev_tl_op  mo_sin_rcnt_tl  mort_acc  \
0                 290.0                    1.0             1.0       1.0
1                 244.0                    1.0             1.0       0.0
2                 265.0                   23.0             3.0       0.0
3                 170.0                   21.0            16.0       5.0
4                 136.0                    7.0             7.0       0.0

   mths_since_recent_bc  mths_since_recent_inq  num_accts_ever_120_pd  \
0                   5.0                    1.0                    4.0
1                  47.0                    NaN                    0.0
2                  24.0                   17.0                    0.0
3                  21.0                    1.0                    1.0
4                   7.0                    7.0                    1.0

   num_actv_bc_tl  num_actv_rev_tl  num_bc_sats  num_bc_tl  num_il_tl  \
0             6.0              9.0          7.0       18.0        2.0
1             1.0              4.0          1.0        2.0        8.0
2             4.0              7.0          5.0       16.0       17.0
3             3.0              5.0          3.0        5.0        1.0
4             3.0              4.0          3.0       12.0       16.0

   num_op_rev_tl  num_rev_accts  num_rev_tl_bal_gt_0  num_sats  \
0           14.0           32.0                  9.0      17.0
1            5.0            9.0                  4.0       6.0
```

```
2            8.0           26.0                   7.0      12.0
3            5.0            7.0                   5.0       8.0
4            5.0           18.0                   4.0      10.0

   num_tl_120dpd_2m  num_tl_30dpd  num_tl_90g_dpd_24m  num_tl_op_past_12m  \
0               0.0           0.0                 0.0                 4.0
1               0.0           0.0                 0.0                 4.0
2               0.0           0.0                 0.0                 3.0
3               0.0           0.0                 0.0                 0.0
4               0.0           0.0                 0.0                 2.0

   pct_tl_nvr_dlq  percent_bc_gt_75  pub_rec_bankruptcies  tax_liens  \
0            83.3              14.3                   0.0        0.0
1           100.0               0.0                   0.0        0.0
2           100.0              60.0                   0.0        0.0
3            76.9             100.0                   0.0        0.0
4            91.4             100.0                   0.0        0.0

   tot_hi_cred_lim  total_bal_ex_mort  total_bc_limit  \
0         179407.0            15030.0         13000.0
1         196500.0           149140.0         10000.0
2          52490.0            38566.0         21100.0
3         368700.0            18007.0          4400.0
4          57073.0            42315.0         15000.0

   total_il_high_credit_limit debt_settlement_flag
0                     11325.0                    N
1                     12000.0                    N
2                     24890.0                    N
3                     18000.0                    N
4                     35573.0                    N
```

dummy coding

[26]:
```
cat_columns = ['term','home_ownership','verification_status',
               'purpose','initial_list_status','debt_settlement_flag']
dummy_df = pd.get_dummies(df6_1[cat_columns])
df6_1 = pd.concat([df6_1, dummy_df], axis=1)
df6_1 = df6_1.drop(cat_columns, axis=1)
df6_1.head()
```

[26]:
```
   loan_amnt  int_rate  installment  emp_length  annual_inc  loan_status  \
0    10400.0      6.99       321.08         8.0     58000.0            1
1    15000.0     12.39       336.64        10.0     78000.0            0
2     9600.0     13.66       326.53        10.0     69000.0            0
3    12800.0     17.14       319.08        10.0    125000.0            0
4    21425.0     15.59       516.36         6.0     63800.0            0
```

```
     dti  delinq_2yrs  fico_range_low  fico_range_high  inq_last_6mths  \
0  14.92          0.0           710.0            714.0             2.0
1  12.03          0.0           750.0            754.0             0.0
2  25.81          0.0           680.0            684.0             0.0
3   8.31          1.0           665.0            669.0             0.0
4  18.49          0.0           685.0            689.0             0.0

   open_acc  pub_rec  revol_bal  revol_util  total_acc  last_fico_range_high  \
0      17.0      0.0     6133.0        31.6       36.0                 564.0
1       6.0      0.0   138008.0        29.0       17.0                 714.0
2      12.0      0.0    16388.0        59.4       44.0                 714.0
3       8.0      0.0     5753.0       100.9       13.0                 704.0
4      10.0      0.0    16374.0        76.2       35.0                 529.0

   last_fico_range_low  collections_12_mths_ex_med  acc_now_delinq  \
0                560.0                         0.0             0.0
1                710.0                         0.0             0.0
2                710.0                         0.0             0.0
3                700.0                         0.0             0.0
4                525.0                         0.0             0.0

   tot_coll_amt  tot_cur_bal  total_rev_hi_lim  acc_open_past_24mths  \
0           0.0     162110.0           19400.0                   7.0
1           0.0     149140.0          184500.0                   5.0
2           0.0      38566.0           27600.0                   8.0
3           0.0     261815.0            5700.0                   2.0
4           0.0      42315.0           21500.0                   4.0

   avg_cur_bal  bc_open_to_buy  bc_util  chargeoff_within_12_mths  \
0       9536.0          7599.0     41.5                       0.0
1      29828.0          9525.0      4.7                       0.0
2       3214.0          6494.0     69.2                       0.0
3      32727.0             0.0    103.2                       0.0
4       4232.0           324.0     97.8                       0.0

   delinq_amnt  mo_sin_old_il_acct  mo_sin_old_rev_tl_op  \
0          0.0                76.0                 290.0
1          0.0               103.0                 244.0
2          0.0               183.0                 265.0
3          0.0                16.0                 170.0
4          0.0               135.0                 136.0

   mo_sin_rcnt_rev_tl_op  mo_sin_rcnt_tl  mort_acc  mths_since_recent_bc  \
0                    1.0             1.0       1.0                   5.0
1                    1.0             1.0       0.0                  47.0
2                   23.0             3.0       0.0                  24.0
```

```
3                  21.0           16.0       5.0                  21.0
4                   7.0            7.0       0.0                   7.0

   mths_since_recent_inq  num_accts_ever_120_pd  num_actv_bc_tl  \
0                    1.0                    4.0             6.0
1                    NaN                    0.0             1.0
2                   17.0                    0.0             4.0
3                    1.0                    1.0             3.0
4                    7.0                    1.0             3.0

   num_actv_rev_tl  num_bc_sats  num_bc_tl  num_il_tl  num_op_rev_tl  \
0              9.0          7.0       18.0        2.0           14.0
1              4.0          1.0        2.0        8.0            5.0
2              7.0          5.0       16.0       17.0            8.0
3              5.0          3.0        5.0        1.0            5.0
4              4.0          3.0       12.0       16.0            5.0

   num_rev_accts  num_rev_tl_bal_gt_0  num_sats  num_tl_120dpd_2m  \
0           32.0                  9.0      17.0               0.0
1            9.0                  4.0       6.0               0.0
2           26.0                  7.0      12.0               0.0
3            7.0                  5.0       8.0               0.0
4           18.0                  4.0      10.0               0.0

   num_tl_30dpd  num_tl_90g_dpd_24m  num_tl_op_past_12m  pct_tl_nvr_dlq  \
0           0.0                 0.0                 4.0            83.3
1           0.0                 0.0                 4.0           100.0
2           0.0                 0.0                 3.0           100.0
3           0.0                 0.0                 0.0            76.9
4           0.0                 0.0                 2.0            91.4

   percent_bc_gt_75  pub_rec_bankruptcies  tax_liens  tot_hi_cred_lim  \
0              14.3                   0.0        0.0         179407.0
1               0.0                   0.0        0.0         196500.0
2              60.0                   0.0        0.0          52490.0
3             100.0                   0.0        0.0         368700.0
4             100.0                   0.0        0.0          57073.0

   total_bal_ex_mort  total_bc_limit  total_il_high_credit_limit  \
0            15030.0         13000.0                     11325.0
1           149140.0         10000.0                     12000.0
2            38566.0         21100.0                     24890.0
3            18007.0          4400.0                     18000.0
4            42315.0         15000.0                     35573.0

   term_ 36 months  term_ 60 months  home_ownership_MORTGAGE  \
0                1                0                        1
```

```
1                0                1                          0
2                1                0                          0
3                0                1                          1
4                0                1                          0

   home_ownership_OWN  home_ownership_RENT  verification_status_Not Verified  \
0                   0                    0                                 1
1                   0                    1                                 0
2                   0                    1                                 0
3                   0                    0                                 0
4                   0                    1                                 0

   verification_status_Source Verified  verification_status_Verified  \
0                                    0                             0
1                                    1                             0
2                                    1                             0
3                                    0                             1
4                                    1                             0

   purpose_car  purpose_credit_card  purpose_debt_consolidation  \
0            0                    1                           0
1            0                    0                           1
2            0                    0                           1
3            1                    0                           0
4            0                    1                           0

   purpose_home_improvement  purpose_house  purpose_major_purchase  \
0                         0              0                       0
1                         0              0                       0
2                         0              0                       0
3                         0              0                       0
4                         0              0                       0

   purpose_medical  purpose_moving  purpose_other  purpose_renewable_energy  \
0                0               0              0                         0
1                0               0              0                         0
2                0               0              0                         0
3                0               0              0                         0
4                0               0              0                         0

   purpose_small_business  purpose_vacation  purpose_wedding  \
0                       0                 0                0
1                       0                 0                0
2                       0                 0                0
3                       0                 0                0
4                       0                 0                0
```

```
   initial_list_status_f  initial_list_status_w  debt_settlement_flag_N  \
0                      0                      1                       1
1                      0                      1                       1
2                      1                      0                       1
3                      0                      1                       1
4                      0                      1                       1

   debt_settlement_flag_Y
0                       0
1                       0
2                       0
3                       0
4                       0
```

0.0.9 round6.2 continuous variables

[27]:
```
float_columns_df = df6_1.select_dtypes(include=['float64'])
print(float_columns_df.iloc[0])
```

```
loan_amnt                   10400.00
int_rate                        6.99
installment                   321.08
emp_length                      8.00
annual_inc                  58000.00
dti                            14.92
delinq_2yrs                     0.00
fico_range_low                710.00
fico_range_high               714.00
inq_last_6mths                  2.00
open_acc                       17.00
pub_rec                         0.00
revol_bal                    6133.00
revol_util                     31.60
total_acc                      36.00
last_fico_range_high          564.00
last_fico_range_low           560.00
collections_12_mths_ex_med      0.00
acc_now_delinq                  0.00
tot_coll_amt                    0.00
tot_cur_bal                162110.00
total_rev_hi_lim            19400.00
acc_open_past_24mths            7.00
avg_cur_bal                  9536.00
bc_open_to_buy               7599.00
bc_util                        41.50
chargeoff_within_12_mths        0.00
delinq_amnt                     0.00
```

```
mo_sin_old_il_acct                76.00
mo_sin_old_rev_tl_op             290.00
mo_sin_rcnt_rev_tl_op              1.00
mo_sin_rcnt_tl                     1.00
mort_acc                           1.00
mths_since_recent_bc               5.00
mths_since_recent_inq              1.00
num_accts_ever_120_pd              4.00
num_actv_bc_tl                     6.00
num_actv_rev_tl                    9.00
num_bc_sats                        7.00
num_bc_tl                         18.00
num_il_tl                          2.00
num_op_rev_tl                     14.00
num_rev_accts                     32.00
num_rev_tl_bal_gt_0                9.00
num_sats                          17.00
num_tl_120dpd_2m                   0.00
num_tl_30dpd                       0.00
num_tl_90g_dpd_24m                 0.00
num_tl_op_past_12m                 4.00
pct_tl_nvr_dlq                    83.30
percent_bc_gt_75                  14.30
pub_rec_bankruptcies               0.00
tax_liens                          0.00
tot_hi_cred_lim               179407.00
total_bal_ex_mort              15030.00
total_bc_limit                 13000.00
total_il_high_credit_limit     11325.00
Name: 0, dtype: float64
```

- continues variables contain different range of data, outliers may be a potential problem, will deal with it later
- all the remaining missing value are numerical variables more specifically, integers. Multivarite imputer are used to impute them.

[28]:
```
null_counts = df6_1.isnull().sum()
print(null_counts[null_counts>0])
```

```
emp_length               11561
mo_sin_old_il_acct        6884
mths_since_recent_inq    21167
num_tl_120dpd_2m          7647
dtype: int64
```

[29]:
```
cols = null_counts[null_counts>0].index
df6_1[cols].sample(10)
```

[29]:

	emp_length	mo_sin_old_il_acct	mths_since_recent_inq	num_tl_120dpd_2m
63380	8.0	94.0	4.0	0.0
216546	10.0	161.0	1.0	0.0
199689	10.0	116.0	7.0	0.0
92398	10.0	231.0	NaN	0.0
1719	10.0	146.0	23.0	0.0
78897	8.0	97.0	12.0	0.0
172228	10.0	113.0	6.0	0.0
87129	7.0	NaN	2.0	0.0
138427	2.0	22.0	8.0	0.0
62134	10.0	219.0	6.0	0.0

[30]:
```
imp = IterativeImputer(max_iter=10, random_state=0)
imp.fit(df6_1)
IterativeImputer(random_state=0)
df_filled = imp.transform(df6_1)
```

[31]:
```
df_filled = pd.DataFrame(df_filled, columns = df6_1.columns)
df_filled[cols] = np.round(df_filled[cols])
df_filled.sample(10)
```

[31]:

	loan_amnt	int_rate	installment	emp_length	annual_inc	loan_status	dti	delinq_2yrs	fico_range_low	fico_range_high	inq_last_6mths	\
53256	27000.0	10.15	873.12	10.0	70000.0	0.0	28.89	0.0	695.0	699.0	1.0	
174325	35000.0	19.99	1300.55	5.0	127000.0	0.0						
51766	24000.0	10.99	521.70	9.0	550006.0	0.0						
154037	13000.0	10.99	425.55	5.0	42000.0	0.0						
177534	24000.0	12.49	539.83	10.0	69000.0	1.0						
180948	7000.0	9.67	224.79	8.0	47000.0	0.0						
164750	8000.0	9.67	256.90	10.0	35000.0	0.0						
200543	10000.0	12.99	336.90	10.0	54000.0	0.0						
163202	4000.0	19.47	147.58	2.0	12000.0	0.0						
76302	20000.0	15.61	699.30	10.0	50000.0	0.0						

```
174325  15.30          0.0           680.0           684.0             0.0
51766    2.47          0.0           705.0           709.0             0.0
154037  25.91          0.0           695.0           699.0             1.0
177534  25.95          0.0           700.0           704.0             0.0
180948   7.05          0.0           710.0           714.0             0.0
164750   7.24          0.0           670.0           674.0             0.0
200543   6.07          0.0           660.0           664.0             2.0
163202  17.20          0.0           680.0           684.0             1.0
76302   27.00          0.0           675.0           679.0             0.0

        open_acc  pub_rec  revol_bal  revol_util  total_acc  \
53256       30.0      0.0    41781.0        51.0       45.0
174325      12.0      0.0    32031.0        86.8       23.0
51766        8.0      0.0    30058.0        92.2       27.0
154037       8.0      0.0    13724.0        42.8       11.0
177534      14.0      0.0    47952.0        63.7       25.0
180948       9.0      1.0     8584.0        37.5       17.0
164750       7.0      0.0     4745.0        59.3       20.0
200543       6.0      1.0     5882.0        76.0       25.0
163202       9.0      0.0     4155.0        43.7       16.0
76302        9.0      0.0    25119.0        75.9       34.0

        last_fico_range_high  last_fico_range_low  collections_12_mths_ex_med  \
53256                  644.0                640.0                         0.0
174325                 534.0                530.0                         0.0
51766                  704.0                700.0                         0.0
154037                 584.0                580.0                         0.0
177534                 559.0                555.0                         0.0
180948                 754.0                750.0                         0.0
164750                 769.0                765.0                         1.0
200543                 694.0                690.0                         0.0
163202                 654.0                650.0                         0.0
76302                  569.0                565.0                         0.0

        acc_now_delinq  tot_coll_amt  tot_cur_bal  total_rev_hi_lim  \
53256              0.0           0.0      71948.0           81145.0
174325             0.0           0.0     315103.0           36900.0
51766              0.0           0.0      47584.0           32600.0
154037             0.0           0.0      24755.0           32100.0
177534             0.0           0.0     208562.0           75300.0
180948             0.0           0.0     111495.0           22900.0
164750             0.0         689.0      54145.0            8000.0
200543             0.0           0.0       6149.0            7750.0
163202             0.0           0.0      11905.0            9500.0
76302              0.0           0.0     161175.0           33100.0

        acc_open_past_24mths  avg_cur_bal  bc_open_to_buy  bc_util  \
```

53256	10.0	2767.0	5338.0	71.0
174325	4.0	28646.0	3556.0	89.7
51766	3.0	5948.0	2542.0	92.2
154037	3.0	3094.0	18376.0	42.8
177534	4.0	16043.0	22648.0	67.9
180948	2.0	12388.0	276.0	86.2
164750	4.0	7735.0	1616.0	73.1
200543	5.0	1024.0	1625.0	96.0
163202	7.0	1323.0	3086.0	26.5
76302	3.0	20147.0	774.0	94.9

	chargeoff_within_12_mths	delinq_amnt	mo_sin_old_il_acct	\
53256	0.0	0.0	199.0	
174325	0.0	0.0	82.0	
51766	0.0	0.0	290.0	
154037	0.0	0.0	85.0	
177534	0.0	0.0	46.0	
180948	0.0	0.0	167.0	
164750	0.0	0.0	85.0	
200543	0.0	0.0	245.0	
163202	0.0	0.0	32.0	
76302	0.0	0.0	108.0	

	mo_sin_old_rev_tl_op	mo_sin_rcnt_rev_tl_op	mo_sin_rcnt_tl	mort_acc	\
53256	272.0	6.0	6.0	2.0	
174325	138.0	21.0	4.0	3.0	
51766	248.0	16.0	2.0	3.0	
154037	105.0	3.0	3.0	0.0	
177534	179.0	17.0	6.0	4.0	
180948	146.0	14.0	14.0	2.0	
164750	118.0	11.0	11.0	0.0	
200543	245.0	3.0	3.0	4.0	
163202	42.0	11.0	4.0	0.0	
76302	158.0	6.0	6.0	4.0	

	mths_since_recent_bc	mths_since_recent_inq	num_accts_ever_120_pd	\
53256	6.0	5.0	0.0	
174325	21.0	8.0	0.0	
51766	16.0	10.0	0.0	
154037	3.0	4.0	0.0	
177534	17.0	14.0	0.0	
180948	15.0	11.0	0.0	
164750	11.0	21.0	0.0	
200543	19.0	3.0	0.0	
163202	11.0	2.0	0.0	
76302	6.0	10.0	0.0	

	num_actv_bc_tl	num_actv_rev_tl	num_bc_sats	num_bc_tl	num_il_tl	\
53256	9.0	24.0	15.0	15.0	5.0	
174325	7.0	8.0	8.0	14.0	5.0	
51766	5.0	5.0	5.0	5.0	14.0	
154037	4.0	4.0	4.0	6.0	5.0	
177534	8.0	8.0	8.0	12.0	4.0	
180948	2.0	6.0	2.0	2.0	2.0	
164750	2.0	3.0	2.0	6.0	12.0	
200543	2.0	5.0	3.0	6.0	5.0	
163202	3.0	5.0	4.0	6.0	4.0	
76302	4.0	6.0	4.0	16.0	4.0	

	num_op_rev_tl	num_rev_accts	num_rev_tl_bal_gt_0	num_sats	\
53256	24.0	39.0	16.0	26.0	
174325	9.0	15.0	8.0	12.0	
51766	5.0	10.0	5.0	8.0	
154037	4.0	6.0	4.0	8.0	
177534	10.0	16.0	8.0	14.0	
180948	8.0	13.0	6.0	9.0	
164750	3.0	8.0	3.0	7.0	
200543	5.0	16.0	5.0	6.0	
163202	6.0	12.0	5.0	9.0	
76302	7.0	26.0	6.0	9.0	

	num_tl_120dpd_2m	num_tl_30dpd	num_tl_90g_dpd_24m	\
53256	-0.0	0.0	0.0	
174325	0.0	0.0	0.0	
51766	0.0	0.0	0.0	
154037	0.0	0.0	0.0	
177534	0.0	0.0	0.0	
180948	0.0	0.0	0.0	
164750	0.0	0.0	0.0	
200543	0.0	0.0	0.0	
163202	0.0	0.0	0.0	
76302	0.0	0.0	0.0	

	num_tl_op_past_12m	pct_tl_nvr_dlq	percent_bc_gt_75	\
53256	5.0	100.0	44.4	
174325	1.0	100.0	85.7	
51766	1.0	100.0	80.0	
154037	1.0	100.0	75.0	
177534	1.0	100.0	25.0	
180948	0.0	100.0	100.0	
164750	2.0	100.0	50.0	
200543	1.0	72.0	100.0	
163202	4.0	100.0	50.0	
76302	1.0	100.0	100.0	

	pub_rec_bankruptcies	tax_liens	tot_hi_cred_lim	total_bal_ex_mort	\
53256	0.0	0.0	152316.0	48482.0	
174325	0.0	0.0	329252.0	65860.0	
51766	0.0	0.0	56059.0	47584.0	
154037	0.0	0.0	55536.0	24755.0	
177534	0.0	0.0	255097.0	58636.0	
180948	1.0	0.0	132080.0	8584.0	
164750	0.0	0.0	57400.0	54145.0	
200543	1.0	0.0	8745.0	6149.0	
163202	0.0	0.0	17250.0	11905.0	
76302	0.0	0.0	175979.0	37422.0	

	total_bc_limit	total_il_high_credit_limit	term_ 36 months	\
53256	48040.0	13971.0	1.0	
174325	34400.0	42070.0	1.0	
51766	32600.0	23459.0	0.0	
154037	32100.0	23436.0	1.0	
177534	70600.0	16914.0	0.0	
180948	2000.0	0.0	1.0	
164750	6000.0	49400.0	1.0	
200543	3250.0	995.0	1.0	
163202	4200.0	7750.0	1.0	
76302	15100.0	16404.0	1.0	

	term_ 60 months	home_ownership_MORTGAGE	home_ownership_OWN	\
53256	0.0	1.0	0.0	
174325	0.0	1.0	0.0	
51766	1.0	0.0	0.0	
154037	0.0	0.0	0.0	
177534	1.0	1.0	0.0	
180948	0.0	1.0	0.0	
164750	0.0	0.0	0.0	
200543	0.0	0.0	0.0	
163202	0.0	0.0	1.0	
76302	0.0	1.0	0.0	

	home_ownership_RENT	verification_status_Not Verified	\
53256	0.0	0.0	
174325	0.0	0.0	
51766	1.0	0.0	
154037	1.0	1.0	
177534	0.0	0.0	
180948	0.0	0.0	
164750	1.0	0.0	
200543	1.0	0.0	
163202	0.0	1.0	

```
76302                   0.0                              0.0

        verification_status_Source Verified  verification_status_Verified  \
53256                                   0.0                           1.0
174325                                  0.0                           1.0
51766                                   1.0                           0.0
154037                                  0.0                           0.0
177534                                  1.0                           0.0
180948                                  0.0                           1.0
164750                                  1.0                           0.0
200543                                  1.0                           0.0
163202                                  0.0                           0.0
76302                                   1.0                           0.0

        purpose_car  purpose_credit_card  purpose_debt_consolidation  \
53256           0.0                  1.0                         0.0
174325          0.0                  0.0                         1.0
51766           0.0                  1.0                         0.0
154037          0.0                  0.0                         1.0
177534          0.0                  1.0                         0.0
180948          0.0                  0.0                         1.0
164750          0.0                  1.0                         0.0
200543          0.0                  1.0                         0.0
163202          0.0                  0.0                         1.0
76302           0.0                  0.0                         1.0

        purpose_home_improvement  purpose_house  purpose_major_purchase  \
53256                        0.0            0.0                     0.0
174325                       0.0            0.0                     0.0
51766                        0.0            0.0                     0.0
154037                       0.0            0.0                     0.0
177534                       0.0            0.0                     0.0
180948                       0.0            0.0                     0.0
164750                       0.0            0.0                     0.0
200543                       0.0            0.0                     0.0
163202                       0.0            0.0                     0.0
76302                        0.0            0.0                     0.0

        purpose_medical  purpose_moving  purpose_other  \
53256               0.0             0.0            0.0
174325              0.0             0.0            0.0
51766               0.0             0.0            0.0
154037              0.0             0.0            0.0
177534              0.0             0.0            0.0
180948              0.0             0.0            0.0
164750              0.0             0.0            0.0
200543              0.0             0.0            0.0
```

```
163202              0.0             0.0            0.0
76302               0.0             0.0            0.0

        purpose_renewable_energy  purpose_small_business  purpose_vacation  \
53256                        0.0                     0.0               0.0
174325                       0.0                     0.0               0.0
51766                        0.0                     0.0               0.0
154037                       0.0                     0.0               0.0
177534                       0.0                     0.0               0.0
180948                       0.0                     0.0               0.0
164750                       0.0                     0.0               0.0
200543                       0.0                     0.0               0.0
163202                       0.0                     0.0               0.0
76302                        0.0                     0.0               0.0

        purpose_wedding  initial_list_status_f  initial_list_status_w  \
53256               0.0                    1.0                    0.0
174325              0.0                    1.0                    0.0
51766               0.0                    0.0                    1.0
154037              0.0                    0.0                    1.0
177534              0.0                    1.0                    0.0
180948              0.0                    0.0                    1.0
164750              0.0                    0.0                    1.0
200543              0.0                    1.0                    0.0
163202              0.0                    1.0                    0.0
76302               0.0                    1.0                    0.0

        debt_settlement_flag_N  debt_settlement_flag_Y
53256                      1.0                     0.0
174325                     1.0                     0.0
51766                      1.0                     0.0
154037                     1.0                     0.0
177534                     0.0                     1.0
180948                     1.0                     0.0
164750                     1.0                     0.0
200543                     1.0                     0.0
163202                     1.0                     0.0
76302                      1.0                     0.0
```

- no missing value remaining

[32]:
```
null_counts = df_filled.isnull().sum()
print(null_counts[null_counts>0])
print(df_filled.shape)
```

```
Series([], dtype: int64)
(232617, 83)
```

0.0.10 save data

[33]:
```
df_filled.to_csv("LendingClub_2011_2014_cleanedData.csv")
```

identify target attr and perform class mapping

[39]:
```
Class_mapping={label:idx for idx,label in enumerate(np.
 ↪unique(df_filled['loan_status']))}
print(Class_mapping)
df_filled['loan_status']=df_filled['loan_status'].map(Class_mapping)
df_filled.head()
```

```
{0: 0, 1: 1}
```

[39]:

	loan_amnt	int_rate	installment	emp_length	annual_inc	loan_status	\
0	10400.0	6.99	321.08	8.0	58000.0	1	
1	15000.0	12.39	336.64	10.0	78000.0	0	
2	9600.0	13.66	326.53	10.0	69000.0	0	
3	12800.0	17.14	319.08	10.0	125000.0	0	
4	21425.0	15.59	516.36	6.0	63800.0	0	

	dti	delinq_2yrs	fico_range_low	fico_range_high	inq_last_6mths	\
0	14.92	0.0	710.0	714.0	2.0	
1	12.03	0.0	750.0	754.0	0.0	
2	25.81	0.0	680.0	684.0	0.0	
3	8.31	1.0	665.0	669.0	0.0	
4	18.49	0.0	685.0	689.0	0.0	

	open_acc	pub_rec	revol_bal	revol_util	total_acc	last_fico_range_high	\
0	17.0	0.0	6133.0	31.6	36.0	564.0	
1	6.0	0.0	138008.0	29.0	17.0	714.0	
2	12.0	0.0	16388.0	59.4	44.0	714.0	
3	8.0	0.0	5753.0	100.9	13.0	704.0	
4	10.0	0.0	16374.0	76.2	35.0	529.0	

	last_fico_range_low	collections_12_mths_ex_med	acc_now_delinq	\
0	560.0	0.0	0.0	
1	710.0	0.0	0.0	
2	710.0	0.0	0.0	
3	700.0	0.0	0.0	
4	525.0	0.0	0.0	

	tot_coll_amt	tot_cur_bal	total_rev_hi_lim	acc_open_past_24mths	\
0	0.0	162110.0	19400.0	7.0	
1	0.0	149140.0	184500.0	5.0	
2	0.0	38566.0	27600.0	8.0	
3	0.0	261815.0	5700.0	2.0	
4	0.0	42315.0	21500.0	4.0	

```
   avg_cur_bal  bc_open_to_buy  bc_util  chargeoff_within_12_mths  \
0       9536.0          7599.0     41.5                       0.0
1      29828.0          9525.0      4.7                       0.0
2       3214.0          6494.0     69.2                       0.0
3      32727.0             0.0    103.2                       0.0
4       4232.0           324.0     97.8                       0.0

   delinq_amnt  mo_sin_old_il_acct  mo_sin_old_rev_tl_op  \
0          0.0                76.0                 290.0
1          0.0               103.0                 244.0
2          0.0               183.0                 265.0
3          0.0                16.0                 170.0
4          0.0               135.0                 136.0

   mo_sin_rcnt_rev_tl_op  mo_sin_rcnt_tl  mort_acc  mths_since_recent_bc  \
0                    1.0             1.0       1.0                   5.0
1                    1.0             1.0       0.0                  47.0
2                   23.0             3.0       0.0                  24.0
3                   21.0            16.0       5.0                  21.0
4                    7.0             7.0       0.0                   7.0

   mths_since_recent_inq  num_accts_ever_120_pd  num_actv_bc_tl  \
0                    1.0                    4.0             6.0
1                    9.0                    0.0             1.0
2                   17.0                    0.0             4.0
3                    1.0                    1.0             3.0
4                    7.0                    1.0             3.0

   num_actv_rev_tl  num_bc_sats  num_bc_tl  num_il_tl  num_op_rev_tl  \
0              9.0          7.0       18.0        2.0           14.0
1              4.0          1.0        2.0        8.0            5.0
2              7.0          5.0       16.0       17.0            8.0
3              5.0          3.0        5.0        1.0            5.0
4              4.0          3.0       12.0       16.0            5.0

   num_rev_accts  num_rev_tl_bal_gt_0  num_sats  num_tl_120dpd_2m  \
0           32.0                  9.0      17.0               0.0
1            9.0                  4.0       6.0               0.0
2           26.0                  7.0      12.0               0.0
3            7.0                  5.0       8.0               0.0
4           18.0                  4.0      10.0               0.0

   num_tl_30dpd  num_tl_90g_dpd_24m  num_tl_op_past_12m  pct_tl_nvr_dlq  \
0           0.0                 0.0                 4.0            83.3
1           0.0                 0.0                 4.0           100.0
2           0.0                 0.0                 3.0           100.0
```

```
3           0.0                 0.0                 0.0           76.9
4           0.0                 0.0                 2.0           91.4

   percent_bc_gt_75  pub_rec_bankruptcies  tax_liens  tot_hi_cred_lim  \
0              14.3                   0.0        0.0         179407.0
1               0.0                   0.0        0.0         196500.0
2              60.0                   0.0        0.0          52490.0
3             100.0                   0.0        0.0         368700.0
4             100.0                   0.0        0.0          57073.0

   total_bal_ex_mort  total_bc_limit  total_il_high_credit_limit  \
0            15030.0         13000.0                     11325.0
1           149140.0         10000.0                     12000.0
2            38566.0         21100.0                     24890.0
3            18007.0          4400.0                     18000.0
4            42315.0         15000.0                     35573.0

   term_ 36 months  term_ 60 months  home_ownership_MORTGAGE  \
0              1.0              0.0                      1.0
1              0.0              1.0                      0.0
2              1.0              0.0                      0.0
3              0.0              1.0                      1.0
4              0.0              1.0                      0.0

   home_ownership_OWN  home_ownership_RENT  verification_status_Not Verified  \
0                 0.0                  0.0                               1.0
1                 0.0                  1.0                               0.0
2                 0.0                  1.0                               0.0
3                 0.0                  0.0                               0.0
4                 0.0                  1.0                               0.0

   verification_status_Source Verified  verification_status_Verified  \
0                                  0.0                          0.0
1                                  1.0                          0.0
2                                  1.0                          0.0
3                                  0.0                          1.0
4                                  1.0                          0.0

   purpose_car  purpose_credit_card  purpose_debt_consolidation  \
0          0.0                  1.0                         0.0
1          0.0                  0.0                         1.0
2          0.0                  0.0                         1.0
3          1.0                  0.0                         0.0
4          0.0                  1.0                         0.0

   purpose_home_improvement  purpose_house  purpose_major_purchase  \
0                       0.0            0.0                     0.0
```

```
1                       0.0           0.0                     0.0
2                       0.0           0.0                     0.0
3                       0.0           0.0                     0.0
4                       0.0           0.0                     0.0

   purpose_medical  purpose_moving  purpose_other  purpose_renewable_energy  \
0              0.0             0.0            0.0                       0.0
1              0.0             0.0            0.0                       0.0
2              0.0             0.0            0.0                       0.0
3              0.0             0.0            0.0                       0.0
4              0.0             0.0            0.0                       0.0

   purpose_small_business  purpose_vacation  purpose_wedding  \
0                     0.0               0.0              0.0
1                     0.0               0.0              0.0
2                     0.0               0.0              0.0
3                     0.0               0.0              0.0
4                     0.0               0.0              0.0

   initial_list_status_f  initial_list_status_w  debt_settlement_flag_N  \
0                    0.0                    1.0                     1.0
1                    0.0                    1.0                     1.0
2                    1.0                    0.0                     1.0
3                    0.0                    1.0                     1.0
4                    0.0                    1.0                     1.0

   debt_settlement_flag_Y
0                     0.0
1                     0.0
2                     0.0
3                     0.0
4                     0.0
```

0.0.11 normalization

normalize all remaining attributes to (0,1) range

[40]:
```
num=df_filled.drop('loan_status', axis=1)
x=num.values
min_max_scaler=preprocessing.MinMaxScaler(feature_range=(0, 1), copy=True)
num_scaled=min_max_scaler.fit_transform(x)
num_norm=pd.DataFrame(num_scaled,columns=num.columns)
num_norm.head()
```

[40]:
```
   loan_amnt  int_rate  installment  emp_length  annual_inc       dti  \
0   0.276471  0.049352     0.214708    0.714286    0.007336  0.373093
1   0.411765  0.318544     0.225929    0.857143    0.010004  0.300825
```

```
2   0.252941  0.381854     0.218638    0.857143    0.008804  0.645411
3   0.347059  0.555334     0.213265    0.857143    0.016273  0.207802
4   0.600735  0.478066     0.355538    0.571429    0.008110  0.462366

   delinq_2yrs  fico_range_low  fico_range_high  inq_last_6mths  open_acc  \
0     0.000000        0.270270         0.268817        0.333333  0.192771
1     0.000000        0.486486         0.483871        0.000000  0.060241
2     0.000000        0.108108         0.107527        0.000000  0.132530
3     0.045455        0.027027         0.026882        0.000000  0.084337
4     0.000000        0.135135         0.134409        0.000000  0.108434

   pub_rec  revol_bal  revol_util  total_acc  last_fico_range_high  \
0      0.0   0.002395    0.086197   0.220779              0.663529
1      0.0   0.053895    0.079105   0.097403              0.840000
2      0.0   0.006400    0.162029   0.272727              0.840000
3      0.0   0.002247    0.275232   0.071429              0.828235
4      0.0   0.006394    0.207856   0.214286              0.622353

   last_fico_range_low  collections_12_mths_ex_med  acc_now_delinq  \
0             0.662722                         0.0             0.0
1             0.840237                         0.0             0.0
2             0.840237                         0.0             0.0
3             0.828402                         0.0             0.0
4             0.621302                         0.0             0.0

   tot_coll_amt  tot_cur_bal  total_rev_hi_lim  acc_open_past_24mths  \
0           0.0     0.042207           0.00192              0.132075
1           0.0     0.038831           0.01843              0.094340
2           0.0     0.010041           0.00274              0.150943
3           0.0     0.068167           0.00055              0.037736
4           0.0     0.011017           0.00213              0.075472

   avg_cur_bal  bc_open_to_buy   bc_util  chargeoff_within_12_mths  \
0     0.019168        0.029199  0.162618                       0.0
1     0.059958        0.036599  0.018417                       0.0
2     0.006461        0.024953  0.271160                       0.0
3     0.065785        0.000000  0.404389                       0.0
4     0.008507        0.001245  0.383229                       0.0

   delinq_amnt  mo_sin_old_il_acct  mo_sin_old_rev_tl_op  \
0          0.0            0.135472              0.341289
1          0.0            0.183601              0.286396
2          0.0            0.326203              0.311456
3          0.0            0.028520              0.198091
4          0.0            0.240642              0.157518

   mo_sin_rcnt_rev_tl_op  mo_sin_rcnt_tl  mort_acc  mths_since_recent_bc  \
```

```
0              0.002688        0.004425  0.027027              0.008117
1              0.002688        0.004425  0.000000              0.076299
2              0.061828        0.013274  0.000000              0.038961
3              0.056452        0.070796  0.135135              0.034091
4              0.018817        0.030973  0.000000              0.011364

   mths_since_recent_inq  num_accts_ever_120_pd  num_actv_bc_tl  \
0                   0.04               0.121212        0.230769
1                   0.36               0.000000        0.038462
2                   0.68               0.000000        0.153846
3                   0.04               0.030303        0.115385
4                   0.28               0.030303        0.115385

   num_actv_rev_tl  num_bc_sats  num_bc_tl  num_il_tl  num_op_rev_tl  \
0         0.236842     0.200000   0.283333   0.013333       0.213115
1         0.105263     0.028571   0.016667   0.053333       0.065574
2         0.184211     0.142857   0.250000   0.113333       0.114754
3         0.131579     0.085714   0.066667   0.006667       0.065574
4         0.105263     0.085714   0.183333   0.106667       0.065574

   num_rev_accts  num_rev_tl_bal_gt_0  num_sats  num_tl_120dpd_2m  \
0       0.291262             0.236842  0.192771               0.0
1       0.067961             0.105263  0.060241               0.0
2       0.233010             0.184211  0.132530               0.0
3       0.048544             0.131579  0.084337               0.0
4       0.155340             0.105263  0.108434               0.0

   num_tl_30dpd  num_tl_90g_dpd_24m  num_tl_op_past_12m  pct_tl_nvr_dlq  \
0           0.0                 0.0            0.153846        0.799520
1           0.0                 0.0            0.153846        1.000000
2           0.0                 0.0            0.115385        1.000000
3           0.0                 0.0            0.000000        0.722689
4           0.0                 0.0            0.076923        0.896759

   percent_bc_gt_75  pub_rec_bankruptcies  tax_liens  tot_hi_cred_lim  \
0             0.143                   0.0        0.0         0.017911
1             0.000                   0.0        0.0         0.019621
2             0.600                   0.0        0.0         0.005219
3             1.000                   0.0        0.0         0.036841
4             1.000                   0.0        0.0         0.005677

   total_bal_ex_mort  total_bc_limit  total_il_high_credit_limit  \
0           0.005590        0.011828                    0.009120
1           0.055465        0.009078                    0.009664
2           0.014343        0.019255                    0.020044
3           0.006697        0.003943                    0.014495
4           0.015737        0.013662                    0.028647
```

```
   term_ 36 months  term_ 60 months  home_ownership_MORTGAGE  \
0              1.0              0.0                      1.0
1              0.0              1.0                      0.0
2              1.0              0.0                      0.0
3              0.0              1.0                      1.0
4              0.0              1.0                      0.0

   home_ownership_OWN  home_ownership_RENT  verification_status_Not Verified  \
0                 0.0                  0.0                               1.0
1                 0.0                  1.0                               0.0
2                 0.0                  1.0                               0.0
3                 0.0                  0.0                               0.0
4                 0.0                  1.0                               0.0

   verification_status_Source Verified  verification_status_Verified  \
0                                  0.0                          0.0
1                                  1.0                          0.0
2                                  1.0                          0.0
3                                  0.0                          1.0
4                                  1.0                          0.0

   purpose_car  purpose_credit_card  purpose_debt_consolidation  \
0          0.0                  1.0                         0.0
1          0.0                  0.0                         1.0
2          0.0                  0.0                         1.0
3          1.0                  0.0                         0.0
4          0.0                  1.0                         0.0

   purpose_home_improvement  purpose_house  purpose_major_purchase  \
0                       0.0            0.0                     0.0
1                       0.0            0.0                     0.0
2                       0.0            0.0                     0.0
3                       0.0            0.0                     0.0
4                       0.0            0.0                     0.0

   purpose_medical  purpose_moving  purpose_other  purpose_renewable_energy  \
0              0.0             0.0            0.0                       0.0
1              0.0             0.0            0.0                       0.0
2              0.0             0.0            0.0                       0.0
3              0.0             0.0            0.0                       0.0
4              0.0             0.0            0.0                       0.0

   purpose_small_business  purpose_vacation  purpose_wedding  \
0                     0.0               0.0              0.0
1                     0.0               0.0              0.0
2                     0.0               0.0              0.0
```

```
3                    0.0                0.0              0.0
4                    0.0                0.0              0.0

   initial_list_status_f  initial_list_status_w  debt_settlement_flag_N  \
0                    0.0                    1.0                     1.0
1                    0.0                    1.0                     1.0
2                    1.0                    0.0                     1.0
3                    0.0                    1.0                     1.0
4                    0.0                    1.0                     1.0

   debt_settlement_flag_Y
0                     0.0
1                     0.0
2                     0.0
3                     0.0
4                     0.0
```

[36]:
```
dfNormalized=pd.merge(df_filled['loan_status'],num_norm,left_index=True,␣
 ↪right_index=True)
dfNormalized.to_csv("loan_normalized",index=False)
dfNormalized.head()
```

[36]:
```
   loan_status  loan_amnt  int_rate  installment  emp_length  annual_inc  \
0            1   0.276471  0.049352     0.214708    0.714286    0.007336
1            0   0.411765  0.318544     0.225929    0.857143    0.010004
2            0   0.252941  0.381854     0.218638    0.857143    0.008804
3            0   0.347059  0.555334     0.213265    0.857143    0.016273
4            0   0.600735  0.478066     0.355538    0.571429    0.008110

        dti  delinq_2yrs  fico_range_low  fico_range_high  inq_last_6mths  \
0  0.373093     0.000000        0.270270         0.268817        0.333333
1  0.300825     0.000000        0.486486         0.483871        0.000000
2  0.645411     0.000000        0.108108         0.107527        0.000000
3  0.207802     0.045455        0.027027         0.026882        0.000000
4  0.462366     0.000000        0.135135         0.134409        0.000000

   open_acc  pub_rec  revol_bal  revol_util  total_acc  last_fico_range_high  \
0  0.192771      0.0   0.002395    0.086197   0.220779              0.663529
1  0.060241      0.0   0.053895    0.079105   0.097403              0.840000
2  0.132530      0.0   0.006400    0.162029   0.272727              0.840000
3  0.084337      0.0   0.002247    0.275232   0.071429              0.828235
4  0.108434      0.0   0.006394    0.207856   0.214286              0.622353

   last_fico_range_low  collections_12_mths_ex_med  acc_now_delinq  \
0             0.662722                         0.0             0.0
1             0.840237                         0.0             0.0
2             0.840237                         0.0             0.0
```

```
3            0.828402                        0.0            0.0
4            0.621302                        0.0            0.0

   tot_coll_amt  tot_cur_bal  total_rev_hi_lim  acc_open_past_24mths  \
0           0.0     0.042207           0.00192              0.132075
1           0.0     0.038831           0.01843              0.094340
2           0.0     0.010041           0.00274              0.150943
3           0.0     0.068167           0.00055              0.037736
4           0.0     0.011017           0.00213              0.075472

   avg_cur_bal  bc_open_to_buy   bc_util  chargeoff_within_12_mths  \
0     0.019168        0.029199  0.162618                       0.0
1     0.059958        0.036599  0.018417                       0.0
2     0.006461        0.024953  0.271160                       0.0
3     0.065785        0.000000  0.404389                       0.0
4     0.008507        0.001245  0.383229                       0.0

   delinq_amnt  mo_sin_old_il_acct  mo_sin_old_rev_tl_op  \
0          0.0            0.135472              0.341289
1          0.0            0.183601              0.286396
2          0.0            0.326203              0.311456
3          0.0            0.028520              0.198091
4          0.0            0.240642              0.157518

   mo_sin_rcnt_rev_tl_op  mo_sin_rcnt_tl  mort_acc  mths_since_recent_bc  \
0               0.002688        0.004425  0.027027              0.008117
1               0.002688        0.004425  0.000000              0.076299
2               0.061828        0.013274  0.000000              0.038961
3               0.056452        0.070796  0.135135              0.034091
4               0.018817        0.030973  0.000000              0.011364

   mths_since_recent_inq  num_accts_ever_120_pd  num_actv_bc_tl  \
0                   0.04               0.121212        0.230769
1                   0.36               0.000000        0.038462
2                   0.68               0.000000        0.153846
3                   0.04               0.030303        0.115385
4                   0.28               0.030303        0.115385

   num_actv_rev_tl  num_bc_sats  num_bc_tl  num_il_tl  num_op_rev_tl  \
0         0.236842     0.200000   0.283333   0.013333       0.213115
1         0.105263     0.028571   0.016667   0.053333       0.065574
2         0.184211     0.142857   0.250000   0.113333       0.114754
3         0.131579     0.085714   0.066667   0.006667       0.065574
4         0.105263     0.085714   0.183333   0.106667       0.065574

   num_rev_accts  num_rev_tl_bal_gt_0  num_sats  num_tl_120dpd_2m  \
0       0.291262             0.236842  0.192771               0.0
```

```
1       0.067961             0.105263  0.060241              0.0
2       0.233010             0.184211  0.132530              0.0
3       0.048544             0.131579  0.084337              0.0
4       0.155340             0.105263  0.108434              0.0

   num_tl_30dpd  num_tl_90g_dpd_24m  num_tl_op_past_12m  pct_tl_nvr_dlq  \
0           0.0                 0.0            0.153846        0.799520
1           0.0                 0.0            0.153846        1.000000
2           0.0                 0.0            0.115385        1.000000
3           0.0                 0.0            0.000000        0.722689
4           0.0                 0.0            0.076923        0.896759

   percent_bc_gt_75  pub_rec_bankruptcies  tax_liens  tot_hi_cred_lim  \
0             0.143                   0.0        0.0         0.017911
1             0.000                   0.0        0.0         0.019621
2             0.600                   0.0        0.0         0.005219
3             1.000                   0.0        0.0         0.036841
4             1.000                   0.0        0.0         0.005677

   total_bal_ex_mort  total_bc_limit  total_il_high_credit_limit  \
0           0.005590        0.011828                    0.009120
1           0.055465        0.009078                    0.009664
2           0.014343        0.019255                    0.020044
3           0.006697        0.003943                    0.014495
4           0.015737        0.013662                    0.028647

   term_ 36 months  term_ 60 months  home_ownership_MORTGAGE  \
0              1.0              0.0                      1.0
1              0.0              1.0                      0.0
2              1.0              0.0                      0.0
3              0.0              1.0                      1.0
4              0.0              1.0                      0.0

   home_ownership_OWN  home_ownership_RENT  verification_status_Not Verified  \
0                 0.0                  0.0                               1.0
1                 0.0                  1.0                               0.0
2                 0.0                  1.0                               0.0
3                 0.0                  0.0                               0.0
4                 0.0                  1.0                               0.0

   verification_status_Source Verified  verification_status_Verified  \
0                                  0.0                          0.0
1                                  1.0                          0.0
2                                  1.0                          0.0
3                                  0.0                          1.0
4                                  1.0                          0.0
```

```
   purpose_car  purpose_credit_card  purpose_debt_consolidation  \
0          0.0                  1.0                         0.0
1          0.0                  0.0                         1.0
2          0.0                  0.0                         1.0
3          1.0                  0.0                         0.0
4          0.0                  1.0                         0.0

   purpose_home_improvement  purpose_house  purpose_major_purchase  \
0                       0.0            0.0                     0.0
1                       0.0            0.0                     0.0
2                       0.0            0.0                     0.0
3                       0.0            0.0                     0.0
4                       0.0            0.0                     0.0

   purpose_medical  purpose_moving  purpose_other  purpose_renewable_energy  \
0              0.0             0.0            0.0                       0.0
1              0.0             0.0            0.0                       0.0
2              0.0             0.0            0.0                       0.0
3              0.0             0.0            0.0                       0.0
4              0.0             0.0            0.0                       0.0

   purpose_small_business  purpose_vacation  purpose_wedding  \
0                     0.0               0.0              0.0
1                     0.0               0.0              0.0
2                     0.0               0.0              0.0
3                     0.0               0.0              0.0
4                     0.0               0.0              0.0

   initial_list_status_f  initial_list_status_w  debt_settlement_flag_N  \
0                    0.0                    1.0                     1.0
1                    0.0                    1.0                     1.0
2                    1.0                    0.0                     1.0
3                    0.0                    1.0                     1.0
4                    0.0                    1.0                     1.0

   debt_settlement_flag_Y
0                     0.0
1                     0.0
2                     0.0
3                     0.0
4                     0.0
```

0.0.12 train test split

split data into 2/3 training and 1/3 testing dataset

[37]:
```
x = dfNormalized.drop('loan_status', axis=1)
y = dfNormalized['loan_status']
x_train, x_test, y_train, y_test = train_test_split( x, y, test_size=1/3,␣
 ↪random_state=0, stratify=y)
print(x_train.shape[0], x_test.shape[0], y_train.shape[0], y_test.shape[0])
```

```
155078 77539 155078 77539
```

save training and test data

[38]:
```
x_train.to_csv("x_train.csv",index=False)
x_test.to_csv("x_test.csv",index=False)
y_train.to_csv("y_train.csv",index=False)
y_test.to_csv("y_test.csv",index=False)
```

[ ]:
```

```