loan_data_analysis

Allen

13/04/2017

1. Load data and examine the variables

```
setwd("/Users/Allen/Desktop/data analytics")
X<-read.csv("loan.csv",header = TRUE, sep = ",")
str(X)
```

```
## 'data.frame':    887379 obs. of  74 variables:
##  $ id                          : int  1077501 1077430 1077175 1076863
1075358 1075269 1069639 1072053 1071795 1071570 ...
##  $ member_id                   : int  1296599 1314167 1313524 1277178
1311748 1311441 1304742 1288686 1306957 1306721 ...
##  $ loan_amnt                   : num  5000 2500 2400 10000 3000 ...
##  $ funded_amnt                 : num  5000 2500 2400 10000 3000 ...
##  $ funded_amnt_inv             : num  4975 2500 2400 10000 3000 ...
##  $ term                        : Factor w/ 2 levels " 36 months"," 60
months": 1 2 1 1 2 1 2 1 2 2 ...
##  $ int_rate                    : num  10.7 15.3 16 13.5 12.7 ...
##  $ installment                 : num  162.9 59.8 84.3 339.3 67.8 ...
##  $ grade                       : Factor w/ 7 levels
"A","B","C","D",..: 2 3 3 3 2 1 3 5 6 2 ...
##  $ sub_grade                   : Factor w/ 35 levels
"A1","A2","A3",..: 7 14 15 11 10 4 15 21 27 10 ...
##  $ emp_title                   : Factor w/ 299273 levels ""," \tAdv
Mtr Proj Fld Rep",..: 1 224800 1 9368 282199 285977 246848 171062 1
256905 ...
##  $ emp_length                  : Factor w/ 12 levels "< 1 year","1
year",..: 3 1 3 3 2 5 10 11 6 1 ...
##  $ home_ownership              : Factor w/ 6 levels
```

```
"ANY","MORTGAGE",..: 6 6 6 6 6 6 6 6 5 6 ...
##  $ annual_inc                 : num  24000 30000 12252 49200 80000
...
##  $ verification_status        : Factor w/ 3 levels "Not
Verified",..: 3 2 1 2 2 2 1 2 2 3 ...
##  $ issue_d                    : Factor w/ 103 levels "Apr-
2008","Apr-2009",..: 22 22 22 22 22 22 22 22 22 22 ...
##  $ loan_status                : Factor w/ 10 levels "Charged
Off",..: 6 1 6 6 2 6 2 6 1 1 ...
##  $ pymnt_plan                 : Factor w/ 2 levels "n","y": 1 1 1 1
1 1 1 1 1 1 ...
##  $ url                        : Factor w/ 887379 levels
"https://www.lendingclub.com/browse/loanDetail.action?loan_id=1000007",
..: 21292 21256 21242 21220 20692 20684 19191 19811 19796 19657 ...
##  $ desc                       : Factor w/ 124471 levels "","\t Loan
for purchase of grand piano. Piano will further diversify an already
profitable business. Monthly budget very high via "| __truncated__,..:
113402 113407 1 113258 113232 1 112347 111631 113230 111646 ...
##  $ purpose                    : Factor w/ 14 levels
"car","credit_card",..: 2 1 12 10 10 14 3 1 12 10 ...
##  $ title                      : Factor w/ 63146 levels
"","\tcredit_card",..: 10496 4975 52500 50874 50267 42595 36948 7263
24371 6112 ...
##  $ zip_code                   : Factor w/ 935 levels
"007xx","008xx",..: 810 296 572 856 909 803 267 839 897 729 ...
##  $ addr_state                 : Factor w/ 51 levels
"AK","AL","AR",..: 4 11 15 5 38 4 28 5 5 44 ...
##  $ dti                        : num  27.65 1 8.72 20 17.94 ...
##  $ delinq_2yrs                : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ earliest_cr_line           : Factor w/ 698 levels "","Apr-
1955",..: 265 43 572 210 276 575 342 287 48 690 ...
##  $ inq_last_6mths             : num  1 5 2 1 0 3 1 2 2 0 ...
##  $ mths_since_last_delinq     : num  NA NA NA 35 38 NA NA NA NA NA
...
##  $ mths_since_last_record     : num  NA NA NA NA NA NA NA NA NA NA
...
##  $ open_acc                   : num  3 3 2 10 15 9 7 4 11 2 ...
##  $ pub_rec                    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ revol_bal                  : num  13648 1687 2956 5598 27783 ...
##  $ revol_util                 : num  83.7 9.4 98.5 21 53.9 28.3 85.6
87.5 32.6 36.5 ...
##  $ total_acc                  : num  9 4 10 37 38 12 11 4 13 3 ...
##  $ initial_list_status        : Factor w/ 2 levels "f","w": 1 1 1 1
1 1 1 1 1 1 ...
##  $ out_prncp                  : num  0 0 0 0 767 ...
##  $ out_prncp_inv              : num  0 0 0 0 767 ...
##  $ total_pymnt                : num  5861 1009 3004 12226 3242 ...
##  $ total_pymnt_inv            : num  5832 1009 3004 12226 3242 ...
##  $ total_rec_prncp            : num  5000 456 2400 10000 2233 ...
##  $ total_rec_int              : num  861 435 604 2209 1009 ...
```

```
##  $ total_rec_late_fee         : num  0 0 0 17 0 ...
##  $ recoveries                 : num  0 117 0 0 0 ...
##  $ collection_recovery_fee    : num  0 1.11 0 0 0 0 0 0 2.09 2.52
...
##  $ last_pymnt_d               : Factor w/ 99 levels "","Apr-
2008",..: 42 7 58 42 43 42 43 42 6 80 ...
##  $ last_pymnt_amnt            : num  171.6 119.7 649.9 357.5 67.8
...
##  $ next_pymnt_d               : Factor w/ 101 levels "","Apr-
2008",..: 1 1 1 1 35 1 35 1 1 1 ...
##  $ last_credit_pull_d         : Factor w/ 104 levels "","Apr-
2009",..: 43 102 43 42 43 104 43 25 14 67 ...
##  $ collections_12_mths_ex_med : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ mths_since_last_major_derog: num  NA NA NA NA NA NA NA NA NA NA
...
##  $ policy_code                : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ application_type           : Factor w/ 2 levels
"INDIVIDUAL","JOINT": 1 1 1 1 1 1 1 1 1 1 ...
##  $ annual_inc_joint           : num  NA NA NA NA NA NA NA NA NA NA
...
##  $ dti_joint                  : num  NA NA NA NA NA NA NA NA NA NA
...
##  $ verification_status_joint  : Factor w/ 4 levels "","Not
Verified",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ acc_now_delinq             : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ tot_coll_amt               : num  NA NA NA NA NA NA NA NA NA NA
...
##  $ tot_cur_bal                : num  NA NA NA NA NA NA NA NA NA NA
...
##  $ open_acc_6m                : num  NA NA NA NA NA NA NA NA NA NA
...
##  $ open_il_6m                 : num  NA NA NA NA NA NA NA NA NA NA
...
##  $ open_il_12m                : num  NA NA NA NA NA NA NA NA NA NA
...
##  $ open_il_24m                : num  NA NA NA NA NA NA NA NA NA NA
...
##  $ mths_since_rcnt_il         : num  NA NA NA NA NA NA NA NA NA NA
...
##  $ total_bal_il               : num  NA NA NA NA NA NA NA NA NA NA
...
##  $ il_util                    : num  NA NA NA NA NA NA NA NA NA NA
...
##  $ open_rv_12m                : num  NA NA NA NA NA NA NA NA NA NA
...
##  $ open_rv_24m                : num  NA NA NA NA NA NA NA NA NA NA
...
##  $ max_bal_bc                 : num  NA NA NA NA NA NA NA NA NA NA
...
##  $ all_util                   : num  NA NA NA NA NA NA NA NA NA NA
```

```
...
##  $ total_rev_hi_lim           : num  NA NA NA NA NA NA NA NA NA NA
...
##  $ inq_fi                     : num  NA NA NA NA NA NA NA NA NA NA
...
##  $ total_cu_tl                : num  NA NA NA NA NA NA NA NA NA NA
...
##  $ inq_last_12m               : num  NA NA NA NA NA NA NA NA NA NA
...
```

2. Transform variables

Date variables: "earliest_cr_line", "last_credit_pull_d"

```r
#earliest_cr_line:
#change to the number of months to 2016-01: the approximate collection
date of the dataset
library(zoo)

## Warning: package 'zoo' was built under R version 3.3.2

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

#create a new variable representing the number of month from the
earliest_cr_line date
date_of_collection = as.Date("2016-01-01")

X$months_from_earliest_cr_line = floor(as.numeric(difftime(
  date_of_collection,
  as.Date(as.yearmon(X$earliest_cr_line, "%b-%Y")),
  units = "weeks"
) / 4))

#last_credit_pull_d
X$months_from_last_credit_pull_d = floor(as.numeric(difftime(
  date_of_collection,
  as.Date(as.yearmon(X$last_credit_pull_d, "%b-%Y")),
  units = "weeks"
) / 4))
```

The date variables are converted to number of months, which may contribute to the model if we treat these variables as numerical input. For "Issued date"", "last_pymnt_date"" and "next_pymnt_date", I will most likely won't include them in the model(I will explain later), so no new variables are created.

"zip_code":

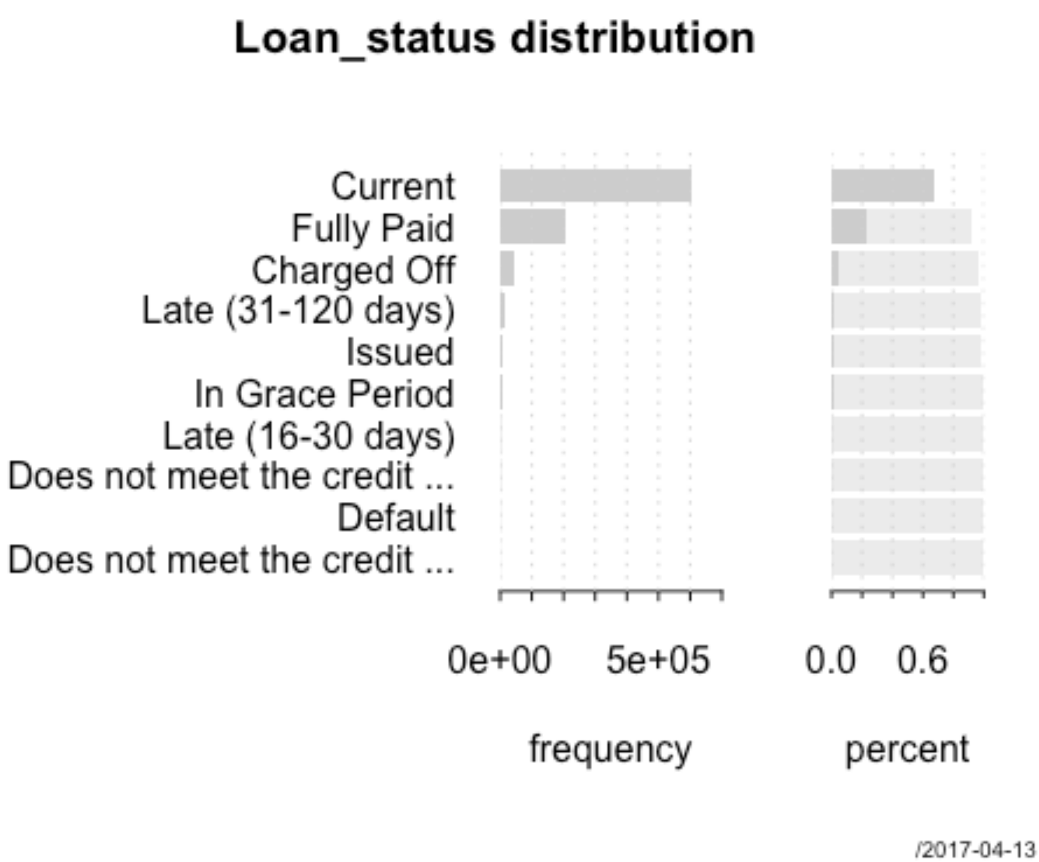
The values all in the format of "number+XX". I will remove the homogenous "XX" and extract the first three letters. This variable is related to state, we may remove it later if it is not important.

```
X$zip_code <- as.factor(gsub("\\D", "", as.character(X$zip_code)))
```

Examine the response variable "loan_status":

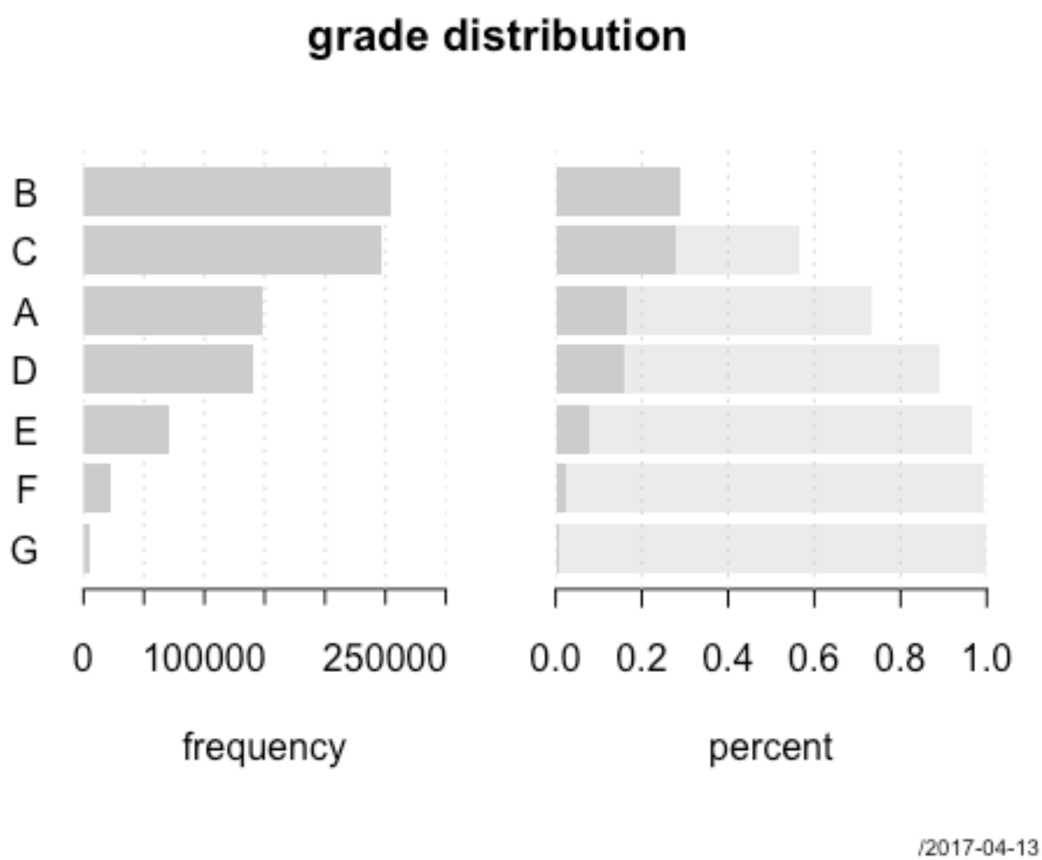
```
summary(X$loan_status)

##                                         Charged Off
##                                               45248
##                                             Current
##                                              601779
##                                             Default
##                                                1219
## Does not meet the credit policy. Status:Charged Off
##                                                 761
##  Does not meet the credit policy. Status:Fully Paid
##                                                1988
##                                          Fully Paid
##                                              207723
##                                     In Grace Period
##                                                6253
##                                              Issued
##                                                8460
##                                   Late (16-30 days)
##                                                2357
##                                  Late (31-120 days)
##                                               11591

library(DescTools)

## Warning: package 'DescTools' was built under R version 3.3.2

Desc(X$loan_status, main = "Loan_status distribution", plotit = 1)

## ---------------------------------------------------------------------
-----
## Loan_status distribution
##
##   length      n    NAs unique levels  dupes
##    9e+05  9e+05      0  1e+01  1e+01      y
##          100.0%   0.0%
##
##                                                      level   freq
perc  cumfreq  cumperc
## 1                                                  Current  6e+05
67.8%    6e+05    67.8%
## 2                                               Fully Paid  2e+05
```

```
23.4%     8e+05     91.2%
## 3                                          Charged Off   5e+04
5.1%     9e+05     96.3%
## 4                                   Late (31-120 days)   1e+04
1.3%     9e+05     97.6%
## 5                                               Issued   8e+03
1.0%     9e+05     98.6%
## 6                                      In Grace Period   6e+03
0.7%     9e+05     99.3%
## 7                                    Late (16-30 days)   2e+03
0.3%     9e+05     99.6%
## 8     Does not meet the credit policy. Status:Fully Paid   2e+03
0.2%     9e+05     99.8%
## 9                                              Default   1e+03
0.1%     9e+05     99.9%
## 10  Does not meet the credit policy. Status:Charged Off   8e+02
0.1%     9e+05    100.0%
```




```
#examine grade
Desc(X$grade, main = "grade distribution", plotit = 1)

## ------------------------------------------------------------------------
-----
```

```
## grade distribution
##
##    length       n     NAs  unique levels   dupes
##     9e+05   9e+05       0   7e+00   7e+00       y
##           100.0%    0.0%
##
##     level    freq    perc  cumfreq   cumperc
## 1       B   3e+05   28.7%    3e+05    28.7%
## 2       C   2e+05   27.7%    5e+05    56.4%
## 3       A   1e+05   16.7%    6e+05    73.1%
## 4       D   1e+05   15.7%    8e+05    88.8%
## 5       E   7e+04    8.0%    9e+05    96.8%
## 6       F   2e+04    2.6%    9e+05    99.4%
## 7       G   5e+03    0.6%    9e+05   100.0%
```




```
#try to find the correlation between loan_status and grade
library(gmodels)
loan_grade<-CrossTable(X$loan_status,X$grade,chisq = TRUE)

##
##
##    Cell Contents
## |-------------------------|
```

```
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  887379
##
##
##                                                      | X$grade
##                                        X$loan_status |         A |
B |          C |          D |          E |          F |          G | Row
Total |
## -----------------------------------------------------|-----------|---
--------|-----------|-----------|-----------|-----------|-----------|--
---------|
##                                          Charged Off |      2617 |
9519 |     12642 |     10486 |      6258 |      2934 |       792 |
45248 |
##                                                      |  3229.193 |
922.334 |     0.887 |  1596.749 |  1951.815 |  2632.582 |   937.017 |
|
##                                                      |     0.058 |
0.210 |     0.279 |     0.232 |     0.138 |     0.065 |     0.018 |
0.051 |
##                                                      |     0.018 |
0.037 |     0.051 |     0.075 |     0.089 |     0.127 |     0.144 |
|
##                                                      |     0.003 |
0.011 |     0.014 |     0.012 |     0.007 |     0.003 |     0.001 |
|
## -----------------------------------------------------|-----------|---
--------|-----------|-----------|-----------|-----------|-----------|--
---------|
##                                              Current |    103322 |
171735 |    171175 |     91984 |     47061 |     13589 |      2913 |
601779 |
##                                                      |    79.031 |
4.474 |   118.461 |    74.034 |    16.439 |   266.206 |   175.990 |
|
##                                                      |     0.172 |
0.285 |     0.284 |     0.153 |     0.078 |     0.023 |     0.005 |
0.678 |
##                                                      |     0.697 |
0.675 |     0.696 |     0.659 |     0.666 |     0.590 |     0.531 |
|
##                                                      |     0.116 |
0.194 |     0.193 |     0.104 |     0.053 |     0.015 |     0.003 |
```

```
|
## ---------------------------------------------------|-----------|---
--------|-----------|-----------|-----------|-----------|-----------|--
---------|
##                                             Default |        47 |
198 |       360 |       312 |       201 |        79 |        22 |
1219 |
##                                                     |   120.437 |
65.778 |     1.467 |    75.510 |   111.084 |    70.794 |    27.729 |
|
##                                                     |     0.039 |
0.162 |     0.295 |     0.256 |     0.165 |     0.065 |     0.018 |
0.001 |
##                                                     |     0.000 |
0.001 |     0.001 |     0.002 |     0.003 |     0.003 |     0.004 |
|
##                                                     |     0.000 |
0.000 |     0.000 |     0.000 |     0.000 |     0.000 |     0.000 |
|
## ---------------------------------------------------|-----------|---
--------|-----------|-----------|-----------|-----------|-----------|--
---------|
## Does not meet the credit policy. Status:Charged Off |         8 |
85 |       148 |       197 |       158 |        93 |        72 |
761 |
##                                                     |   111.599 |
81.384 |    18.732 |    49.972 |   156.343 |   271.381 |   961.983 |
|
##                                                     |     0.011 |
0.112 |     0.194 |     0.259 |     0.208 |     0.122 |     0.095 |
0.001 |
##                                                     |     0.000 |
0.000 |     0.001 |     0.001 |     0.002 |     0.004 |     0.013 |
|
##                                                     |     0.000 |
0.000 |     0.000 |     0.000 |     0.000 |     0.000 |     0.000 |
|
## ---------------------------------------------------|-----------|---
--------|-----------|-----------|-----------|-----------|-----------|--
---------|
##  Does not meet the credit policy. Status:Fully Paid |        90 |
269 |       481 |       494 |       378 |       154 |       122 |
1988 |
##                                                     |   176.414 |
159.133 |     8.846 |   105.240 |   304.442 |   202.975 |   978.670 |
|
##                                                     |     0.045 |
0.135 |     0.242 |     0.248 |     0.190 |     0.077 |     0.061 |
0.002 |
##                                                     |     0.001 |
```

```
0.001 |     0.002 |     0.004 |     0.005 |     0.007 |     0.022 |
|
##                                                      |     0.000 |
0.000 |     0.001 |     0.001 |     0.000 |     0.000 |     0.000 |
|
## -----------------------------------------------------|-----------|---
--------|-----------|-----------|-----------|-----------|-----------|--
---------|
##                                           Fully Paid |     39679 |
66546 |     52678 |     30020 |     12928 |      4726 |      1146 |
207723 |
##                                                      |   716.881 |
813.692 |   412.835 |   214.148 |   793.091 |    82.899 |    15.015 |
|
##                                                      |     0.191 |
0.320 |     0.254 |     0.145 |     0.062 |     0.023 |     0.006 |
0.234 |
##                                                      |     0.268 |
0.261 |     0.214 |     0.215 |     0.183 |     0.205 |     0.209 |
|
##                                                      |     0.045 |
0.075 |     0.059 |     0.034 |     0.015 |     0.005 |     0.001 |
|
## -----------------------------------------------------|-----------|---
--------|-----------|-----------|-----------|-----------|-----------|--
---------|
##                                      In Grace Period |       365 |
1240 |      1887 |      1405 |       908 |       354 |        94 |
6253 |
##                                                      |   441.891 |
170.873 |    13.782 |   180.855 |   337.017 |   226.066 |    79.125 |
|
##                                                      |     0.058 |
0.198 |     0.302 |     0.225 |     0.145 |     0.057 |     0.015 |
0.007 |
##                                                      |     0.002 |
0.005 |     0.008 |     0.010 |     0.013 |     0.015 |     0.017 |
|
##                                                      |     0.000 |
0.001 |     0.002 |     0.002 |     0.001 |     0.000 |     0.000 |
|
## -----------------------------------------------------|-----------|---
--------|-----------|-----------|-----------|-----------|-----------|--
---------|
##                                               Issued |      1448 |
2529 |      2472 |      1185 |       593 |       194 |        39 |
8460 |
##                                                      |     0.871 |
4.316 |     6.995 |    15.881 |     9.752 |     3.009 |     3.396 |
|
```

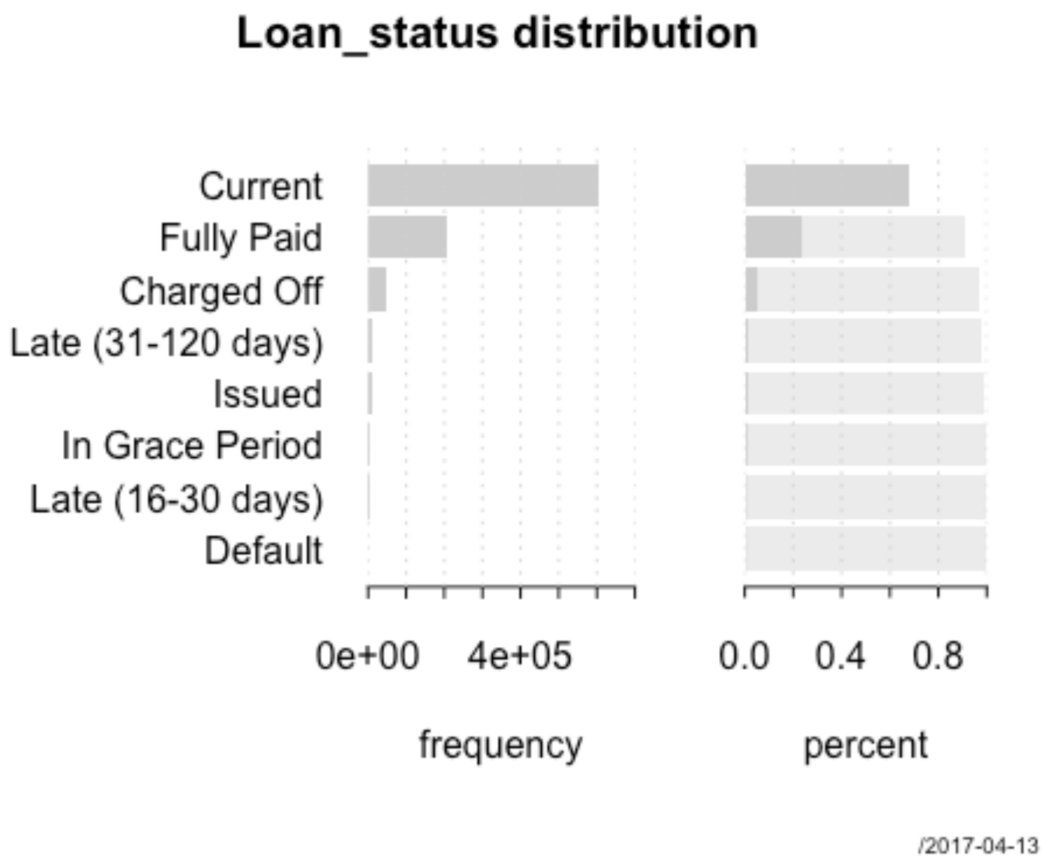
```
##                                                       |     0.171 |
0.299 |     0.292 |     0.140 |     0.070 |     0.023 |     0.005 |
0.010 |
##                                                       |     0.010 |
0.010 |     0.010 |     0.008 |     0.008 |     0.008 |     0.007 |
|
##                                                       |     0.002 |
0.003 |     0.003 |     0.001 |     0.001 |     0.000 |     0.000 |
|
## -----------------------------------------------------|-----------|---
--------|-----------|-----------|-----------|-----------|-----------|--
---------|
##                                   Late (16-30 days) |       134 |
410 |       678 |       569 |       368 |       155 |        43 |
2357 |
##                                                       |   171.260 |
104.719 |     0.954 |   106.155 |   172.901 |   143.693 |    55.401 |
|
##                                                       |     0.057 |
0.174 |     0.288 |     0.241 |     0.156 |     0.066 |     0.018 |
0.003 |
##                                                       |     0.001 |
0.002 |     0.003 |     0.004 |     0.005 |     0.007 |     0.008 |
|
##                                                       |     0.000 |
0.000 |     0.001 |     0.001 |     0.000 |     0.000 |     0.000 |
|
## -----------------------------------------------------|-----------|---
--------|-----------|-----------|-----------|-----------|-----------|--
---------|
##                                  Late (31-120 days) |       492 |
2004 |      3339 |      2890 |      1852 |       768 |       246 |
11591 |
##                                                       |  1076.868 |
524.667 |     5.067 |   624.958 |   933.367 |   724.392 |   423.742 |
|
##                                                       |     0.042 |
0.173 |     0.288 |     0.249 |     0.160 |     0.066 |     0.021 |
0.013 |
##                                                       |     0.003 |
0.008 |     0.014 |     0.021 |     0.026 |     0.033 |     0.045 |
|
##                                                       |     0.001 |
0.002 |     0.004 |     0.003 |     0.002 |     0.001 |     0.000 |
|
## -----------------------------------------------------|-----------|---
--------|-----------|-----------|-----------|-----------|-----------|--
---------|
##                                        Column Total |    148202 |
254535 |    245860 |    139542 |     70705 |     23046 |      5489 |
```

```
887379 |
##                                                        |     0.167 |
0.287 |     0.277 |     0.157 |     0.080 |     0.026 |     0.006 |
|
## ----------------------------------------------------|-----------|---
--------|-----------|-----------|-----------|-----------|-----------|--
---------|
##
##
## Statistics for All Table Factors
##
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 =  25675.66     d.f. =  54     p =  0
##
##
##
```

The code returns a warning message which indicates that Chi-squared approximation may be incorrect. I suspect that it may be because some data have small counts, so I tried to combine data.

```
#Here I assume that these two status are the same as "Charged Off" and
"Fully Paid" respectively
X$loan_status[X$loan_status == 'Does not meet the credit policy.
Status:Charged Off'] <-
  'Charged Off'
X$loan_status[X$loan_status == 'Does not meet the credit policy.
Status:Fully Paid'] <-
  'Fully Paid'

#drop these two factors
X$loan_status=factor(X$loan_status)

#now examine "loan_status" again
Desc(X$loan_status, main = "Loan_status distribution", plotit = 1)

## ---------------------------------------------------------------------
-----
## Loan_status distribution
##
##   length      n    NAs unique levels  dupes
##    9e+05  9e+05      0  8e+00  8e+00      y
##          100.0%   0.0%
##
##                 level   freq   perc  cumfreq  cumperc
## 1             Current  6e+05  67.8%    6e+05    67.8%
## 2          Fully Paid  2e+05  23.6%    8e+05    91.4%
## 3         Charged Off  5e+04   5.2%    9e+05    96.6%
```

```
## 4  Late (31-120 days)  1e+04   1.3%    9e+05    97.9%
## 5              Issued  8e+03   1.0%    9e+05    98.9%
## 6     In Grace Period  6e+03   0.7%    9e+05    99.6%
## 7   Late (16-30 days)  2e+03   0.3%    9e+05    99.9%
## 8             Default  1e+03   0.1%    9e+05   100.0%
```




```
loan_grade<-CrossTable(X$loan_status,X$grade,chisq = TRUE)

## 
## 
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
## 
## 
## Total Observations in Table:  887379 
## 
## 
```

```
##                    | X$grade 
##      X$loan_status |         A |         B |         C |         D |
E |         F |         G | Row Total | 
## -------------------|-----------|-----------|-----------|-----------
|-----------|-----------|-----------|-----------|
##        Charged Off |      2625 |      9604 |     12790 |     10683 |
6416 |      3027 |       864 |     46009 | 
##                    |  3330.755 |   978.311 |     0.142 |  1643.220 |
2063.026 |  2809.132 |  1179.609 |           | 
##                    |     0.057 |     0.209 |     0.278 |     0.232 |
0.139 |     0.066 |     0.019 |     0.052 | 
##                    |     0.018 |     0.038 |     0.052 |     0.077 |
0.091 |     0.131 |     0.157 |           | 
##                    |     0.003 |     0.011 |     0.014 |     0.012 |
0.007 |     0.003 |     0.001 |           | 
## -------------------|-----------|-----------|-----------|-----------
|-----------|-----------|-----------|-----------|
##            Current |    103322 |    171735 |    171175 |     91984 |
47061 |     13589 |      2913 |    601779 | 
##                    |    79.031 |     4.474 |   118.461 |    74.034 |
16.439 |   266.206 |   175.990 |           | 
##                    |     0.172 |     0.285 |     0.284 |     0.153 |
0.078 |     0.023 |     0.005 |     0.678 | 
##                    |     0.697 |     0.675 |     0.696 |     0.659 |
0.666 |     0.590 |     0.531 |           | 
##                    |     0.116 |     0.194 |     0.193 |     0.104 |
0.053 |     0.015 |     0.003 |           | 
## -------------------|-----------|-----------|-----------|-----------
|-----------|-----------|-----------|-----------|
##            Default |        47 |       198 |       360 |       312 |
201 |        79 |        22 |      1219 | 
##                    |   120.437 |    65.778 |     1.467 |    75.510 |
111.084 |    70.794 |    27.729 |           | 
##                    |     0.039 |     0.162 |     0.295 |     0.256 |
0.165 |     0.065 |     0.018 |     0.001 | 
##                    |     0.000 |     0.001 |     0.001 |     0.002 |
0.003 |     0.003 |     0.004 |           | 
##                    |     0.000 |     0.000 |     0.000 |     0.000 |
0.000 |     0.000 |     0.000 |           | 
## -------------------|-----------|-----------|-----------|-----------
|-----------|-----------|-----------|-----------|
##         Fully Paid |     39769 |     66815 |     53159 |     30514 |
13306 |      4880 |      1268 |    209711 | 
##                    |   642.837 |   737.749 |   420.716 |   184.022 |
693.229 |    58.898 |     0.657 |           | 
##                    |     0.190 |     0.319 |     0.253 |     0.146 |
0.063 |     0.023 |     0.006 |     0.236 | 
##                    |     0.268 |     0.262 |     0.216 |     0.219 |
0.188 |     0.212 |     0.231 |           | 
##                    |     0.045 |     0.075 |     0.060 |     0.034 |
```

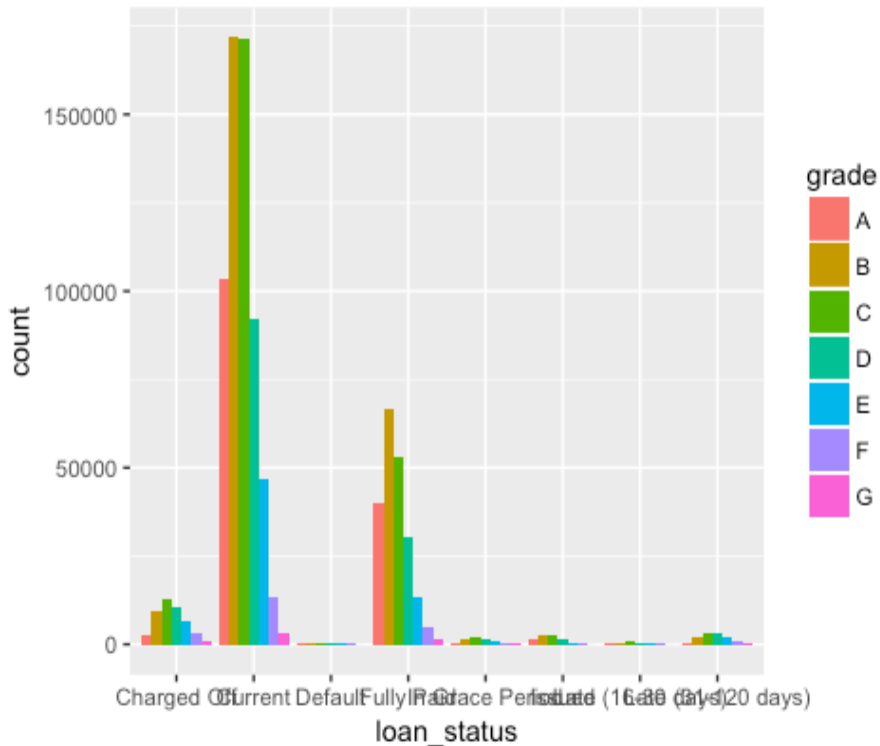
```
0.015 |     0.005 |     0.001 |           |
## -------------------|-----------|-----------|-----------|-----------
|-----------|-----------|-----------|-----------|
##    In Grace Period |       365 |      1240 |      1887 |      1405 |
908 |       354 |        94 |      6253 |
##                    |   441.891 |   170.873 |    13.782 |   180.855 |
337.017 |   226.066 |    79.125 |           |
##                    |     0.058 |     0.198 |     0.302 |     0.225 |
0.145 |     0.057 |     0.015 |     0.007 |
##                    |     0.002 |     0.005 |     0.008 |     0.010 |
0.013 |     0.015 |     0.017 |           |
##                    |     0.000 |     0.001 |     0.002 |     0.002 |
0.001 |     0.000 |     0.000 |           |
## -------------------|-----------|-----------|-----------|-----------
|-----------|-----------|-----------|-----------|
##             Issued |      1448 |      2529 |      2472 |      1185 |
593 |       194 |        39 |      8460 |
##                    |     0.871 |     4.316 |     6.995 |    15.881 |
9.752 |     3.009 |     3.396 |           |
##                    |     0.171 |     0.299 |     0.292 |     0.140 |
0.070 |     0.023 |     0.005 |     0.010 |
##                    |     0.010 |     0.010 |     0.010 |     0.008 |
0.008 |     0.008 |     0.007 |           |
##                    |     0.002 |     0.003 |     0.003 |     0.001 |
0.001 |     0.000 |     0.000 |           |
## -------------------|-----------|-----------|-----------|-----------
|-----------|-----------|-----------|-----------|
##  Late (16-30 days) |       134 |       410 |       678 |       569 |
368 |       155 |        43 |      2357 |
##                    |   171.260 |   104.719 |     0.954 |   106.155 |
172.901 |   143.693 |    55.401 |           |
##                    |     0.057 |     0.174 |     0.288 |     0.241 |
0.156 |     0.066 |     0.018 |     0.003 |
##                    |     0.001 |     0.002 |     0.003 |     0.004 |
0.005 |     0.007 |     0.008 |           |
##                    |     0.000 |     0.000 |     0.001 |     0.001 |
0.000 |     0.000 |     0.000 |           |
## -------------------|-----------|-----------|-----------|-----------
|-----------|-----------|-----------|-----------|
## Late (31-120 days) |       492 |      2004 |      3339 |      2890 |
1852 |       768 |       246 |     11591 |
##                    |  1076.868 |   524.667 |     5.067 |   624.958 |
933.367 |   724.392 |   423.742 |           |
##                    |     0.042 |     0.173 |     0.288 |     0.249 |
0.160 |     0.066 |     0.021 |     0.013 |
##                    |     0.003 |     0.008 |     0.014 |     0.021 |
0.026 |     0.033 |     0.045 |           |
##                    |     0.001 |     0.002 |     0.004 |     0.003 |
0.002 |     0.001 |     0.000 |           |
## -------------------|-----------|-----------|-----------|-----------
```

```
|-----------|-----------|-----------|-----------|
##       Column Total |    148202 |    254535 |    245860 |    139542 |
70705 |     23046 |      5489 |    887379 |
##                    |     0.167 |     0.287 |     0.277 |     0.157 |
0.080 |     0.026 |     0.006 |           |
## -------------------|-----------|-----------|-----------|-----------
|-----------|-----------|-----------|-----------|
##
##
## Statistics for All Table Factors
##
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 =  22511.71     d.f. =  42     p =  0
##
##
##
```

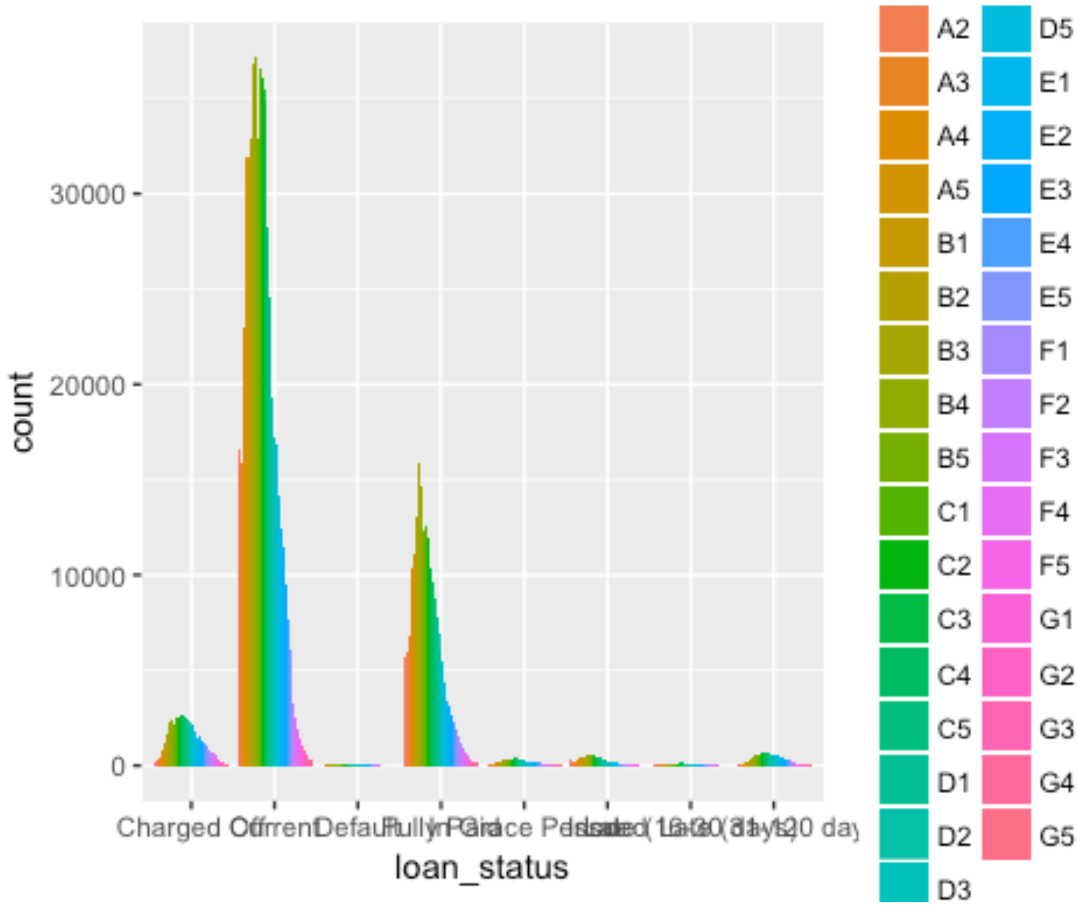
p value = 0, which does not correponds to our expectation. I realize that chi-square test may not be a good way to explore the correaltion between the variables, so I will use visualisation instead.

3. Variable visualisation

```
#loan_status with grade
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.3.2

ggplot(X, aes(loan_status, ..count..)) + geom_bar(aes(fill = grade),
position = "dodge")
```

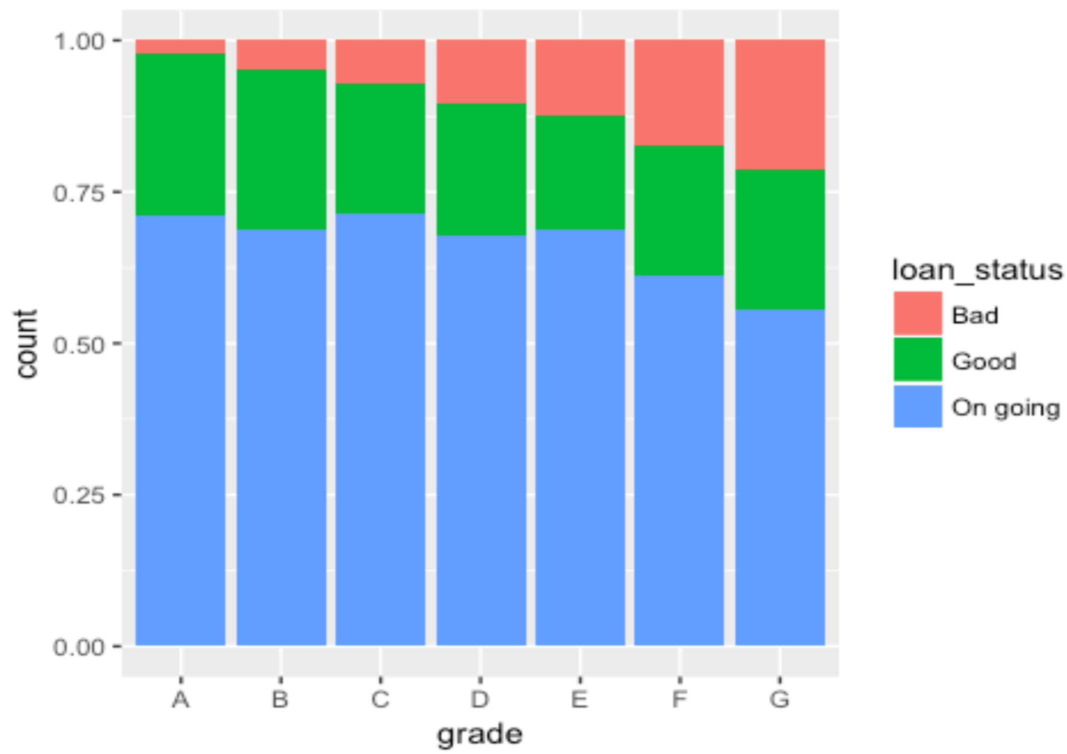



```
#loan_status with sub_grade
ggplot(X, aes(loan_status, ..count..)) + geom_bar(aes(fill =
sub_grade), position = "dodge")
```

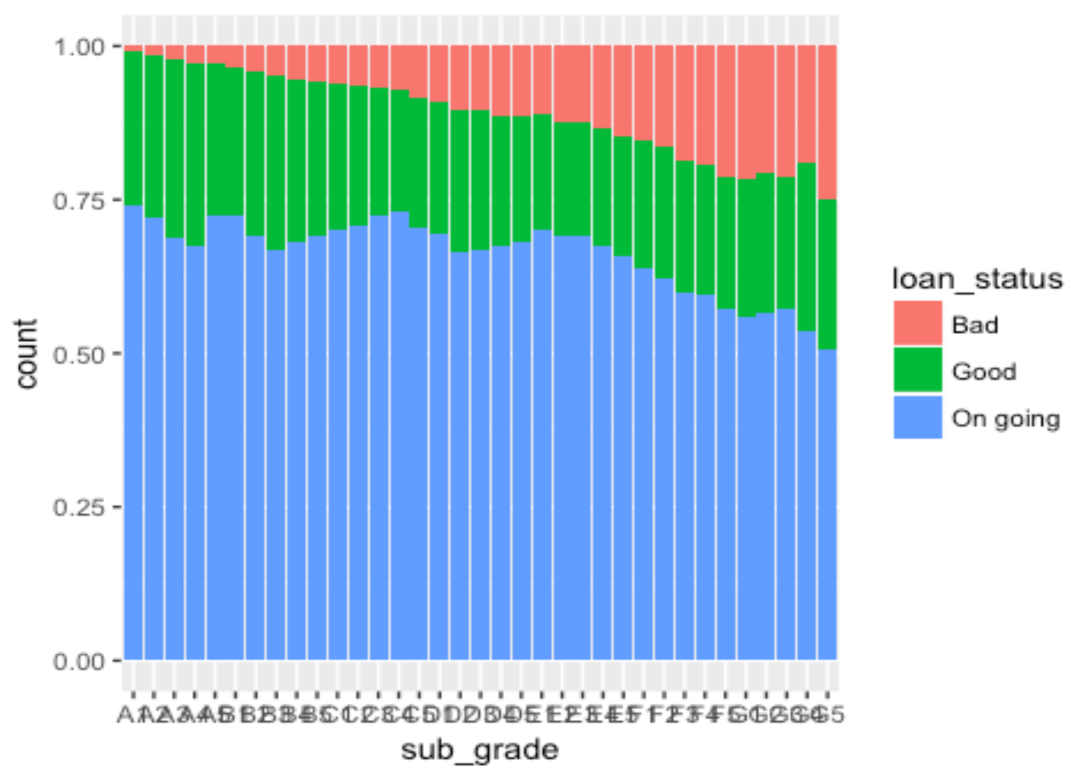


The result shows that both grade and sub_grade affects the possibility of a loan being in the charged off status. However, it is difficult to visualize since there are too many status, so I decided to group "loan_status" into "ongoing"", "paid"" and "bad_status".

```
on_going = c("Current","Issued","In Grace Period")
paid = ("Fully Paid")
X$loan_status = ifelse(X$loan_status %in% paid,"Good",
                              ifelse(X$loan_status %in% on_going,"On
going","Bad"))
#remove unwanted levels
X$loan_status = factor(X$loan_status)

#visualize status by grade and sub_grade again
ggplot(X,aes(grade,fill=loan_status))+geom_bar(position = "fill")
```



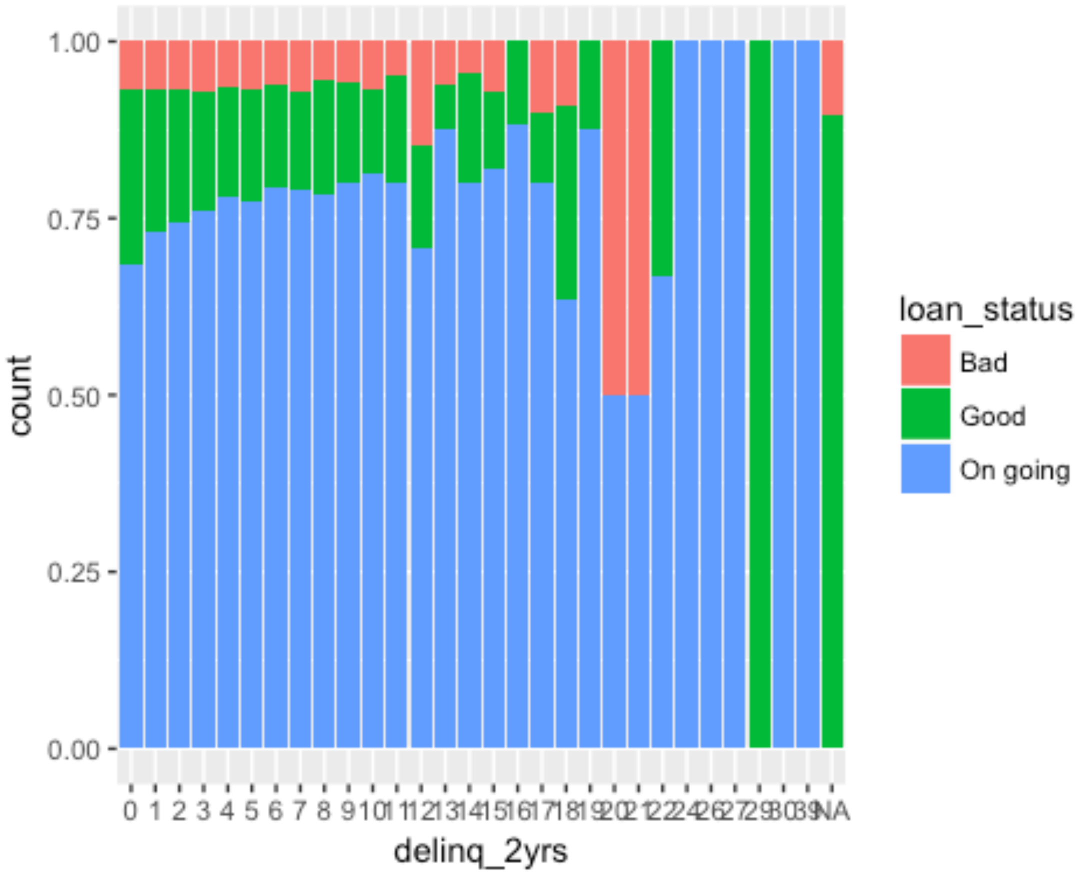

```
ggplot(X,aes(sub_grade,fill=loan_status))+geom_bar(position = "fill")
```



The plot clearly shows that bad status increases when grading increase alphabetically. We then use the similar method to examine other variables.

delinq_2yrs:

```
X$delinq_2yrs<-as.factor(X$delinq_2yrs)
ggplot(X,aes(delinq_2yrs,fill=loan_status))+geom_bar(position = "fill")
```

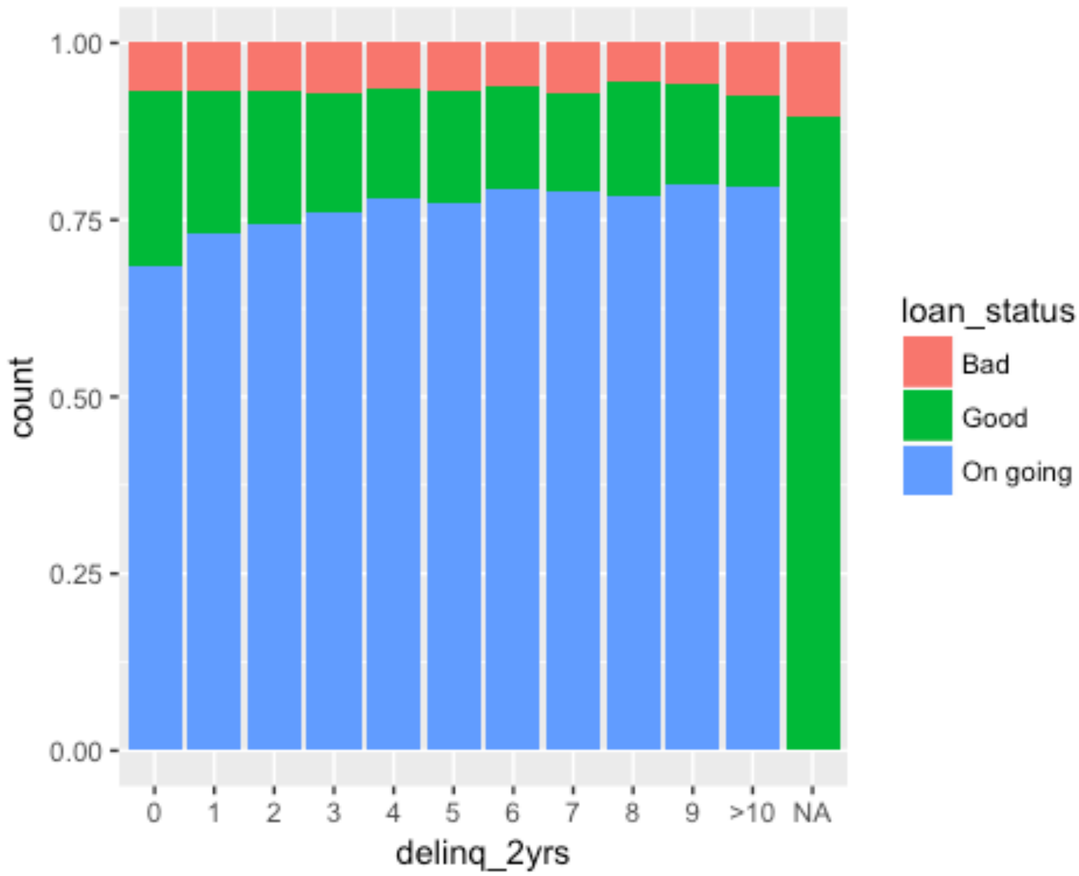


from the plot we can see large proportion of bad status for 21 and 22.

```
summary(X$delinq_2yrs)

##      0      1      2      3      4      5      6      7      8
9
## 716961 113224  33551  11977   5327   2711   1471    784    461
284
##     10     11     12     13     14     15     16     17     18
19
##    192    121     89     64     45     28     17     10     11
8
##     20     21     22     24     26     27     29     30     39
NA's
##      2      2      3      1      2      1      1      1      1
29
```

the summary shows that there are too many classes, so I group the small-count classes into one class. Sum all >10 to one class.

```r
levels(X$delinq_2yrs)<-c(levels(X$delinq_2yrs),">10")
X$delinq_2yrs[as.numeric(X$delinq_2yrs) > 10]<- '>10'

#remove unwanted levels
X$delinq_2yrs=factor(X$delinq_2yrs)

ggplot(X,aes(delinq_2yrs,fill=loan_status))+geom_bar(position = "fill")
```




```r
loan_grade<-CrossTable(X$loan_status,X$delinq_2yrs,chisq = TRUE)
```

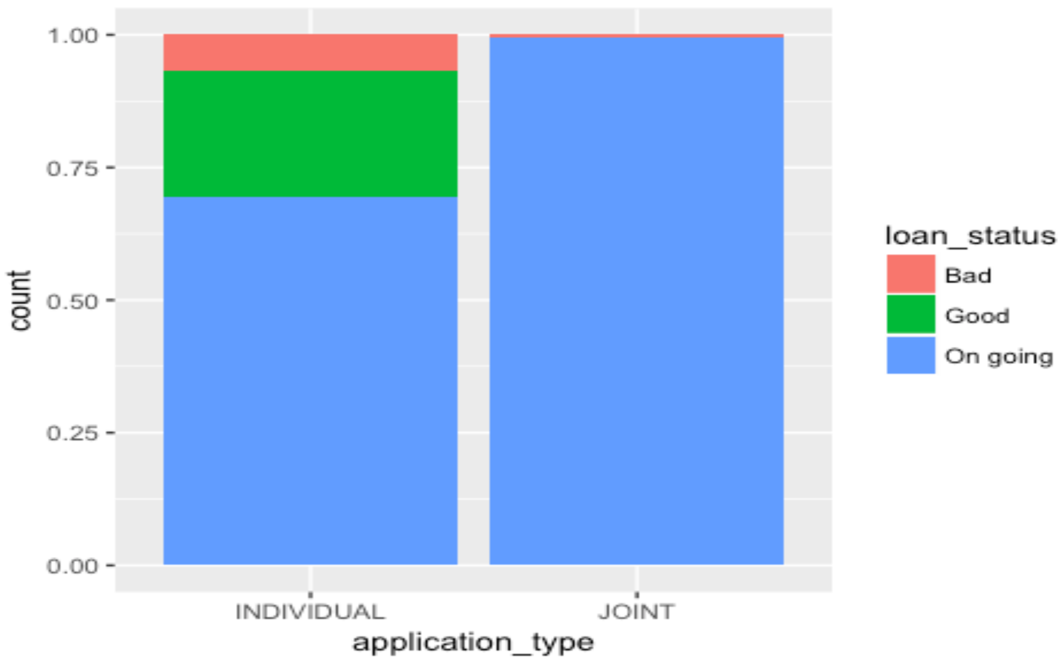
```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  887350
```

```
##
##
##                | X$delinq_2yrs
## X$loan_status |         0 |         1 |         2 |         3 |
4 |         5 |         6 |         7 |         8 |         9 |
>10 | Row Total |
## --------------|-----------|-----------|-----------|-----------|-----
------|-----------|-----------|-----------|-----------|-----------|----
-------|-----------|
##           Bad |     49733 |      7518 |      2317 |       838 |
348 |       187 |        90 |        57 |        25 |        16 |
44 |     61173 |
##               |     1.900 |    10.593 |     0.007 |     0.184 |
1.008 |     0.000 |     1.284 |     0.161 |     1.447 |     0.654 |
0.177 |           |
##               |     0.813 |     0.123 |     0.038 |     0.014 |
0.006 |     0.003 |     0.001 |     0.001 |     0.000 |     0.000 |
0.001 |     0.069 |
##               |     0.069 |     0.066 |     0.069 |     0.070 |
0.065 |     0.069 |     0.061 |     0.073 |     0.054 |     0.056 |
0.073 |           |
##               |     0.056 |     0.008 |     0.003 |     0.001 |
0.000 |     0.000 |     0.000 |     0.000 |     0.000 |     0.000 |
0.000 |           |
## --------------|-----------|-----------|-----------|-----------|-----
------|-----------|-----------|-----------|-----------|-----------|----
-------|-----------|
##          Good |    176458 |     23133 |      6288 |      2038 |
830 |       425 |       212 |       108 |        75 |        41 |
77 |    209685 |
##               |   292.263 |   490.426 |   339.350 |   221.755 |
146.065 |    72.575 |    52.901 |    32.222 |    10.572 |    10.159 |
29.434 |           |
##               |     0.842 |     0.110 |     0.030 |     0.010 |
0.004 |     0.002 |     0.001 |     0.001 |     0.000 |     0.000 |
0.000 |     0.236 |
##               |     0.246 |     0.204 |     0.187 |     0.170 |
0.156 |     0.157 |     0.144 |     0.138 |     0.163 |     0.144 |
0.129 |           |
##               |     0.199 |     0.026 |     0.007 |     0.002 |
0.001 |     0.000 |     0.000 |     0.000 |     0.000 |     0.000 |
0.000 |           |
## --------------|-----------|-----------|-----------|-----------|-----
------|-----------|-----------|-----------|-----------|-----------|----
-------|-----------|
##      On going |    490770 |     82573 |     24946 |      9101 |
4149 |      2099 |      1169 |       619 |       361 |       227 |
478 |    616492 |
##               |   108.253 |   194.340 |   114.855 |    73.097 |
54.238 |    24.660 |    21.148 |    10.138 |     5.176 |     4.467 |
```

```
9.190 |           |
##               |     0.796 |     0.134 |     0.040 |     0.015 |
0.007 |     0.003 |     0.002 |     0.001 |     0.001 |     0.000 |
0.001 |     0.695 |
##               |     0.685 |     0.729 |     0.744 |     0.760 |
0.779 |     0.774 |     0.795 |     0.790 |     0.783 |     0.799 |
0.798 |           |
##               |     0.553 |     0.093 |     0.028 |     0.010 |
0.005 |     0.002 |     0.001 |     0.001 |     0.000 |     0.000 |
0.001 |           |
## --------------|-----------|-----------|-----------|-----------|-----
------|-----------|-----------|-----------|-----------|-----------|----
-------|-----------|
##  Column Total |    716961 |    113224 |     33551 |     11977 |
5327 |      2711 |      1471 |       784 |       461 |       284 |
599 |    887350 |
##               |     0.808 |     0.128 |     0.038 |     0.013 |
0.006 |     0.003 |     0.002 |     0.001 |     0.001 |     0.000 |
0.001 |           |
## --------------|-----------|-----------|-----------|-----------|-----
------|-----------|-----------|-----------|-----------|-----------|----
-------|-----------|
##
##
## Statistics for All Table Factors
##
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 =  2334.698     d.f. =  20     p =  0
##
##
##
```

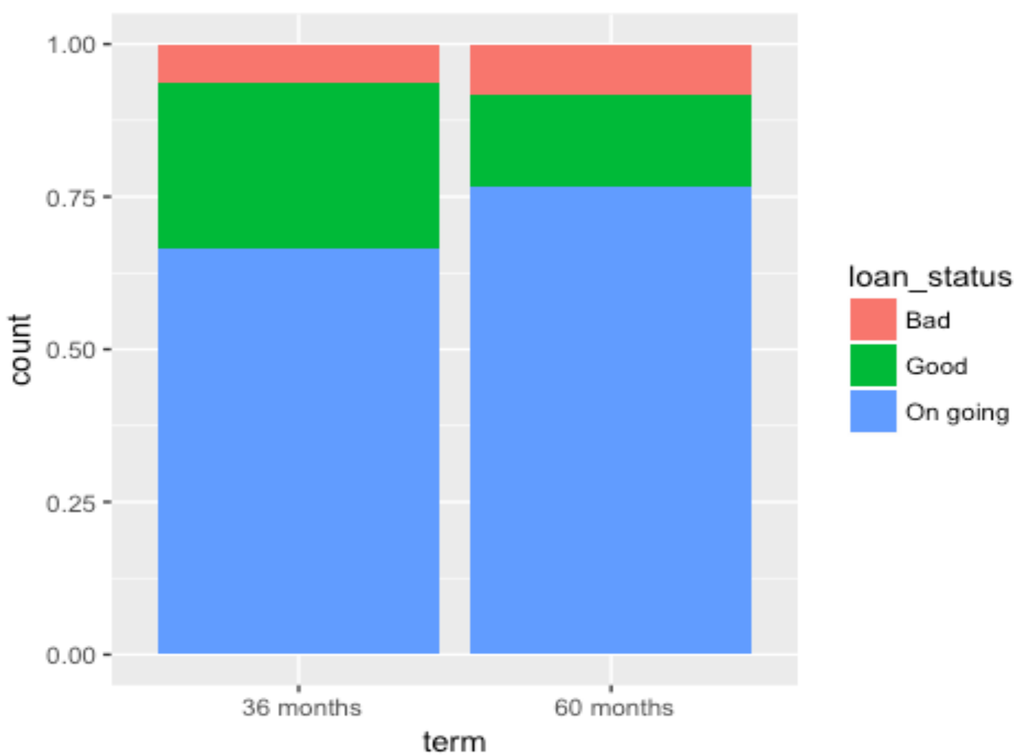
both the ggplot and the chisq test shows that it is not an important feature. However, we only only exclude this from the final model after exmaming its importance during the stage of model building.

```
#examine application type: individual and joint
ggplot(X,aes(application_type,fill=loan_status))+geom_bar(position =
"fill")
```



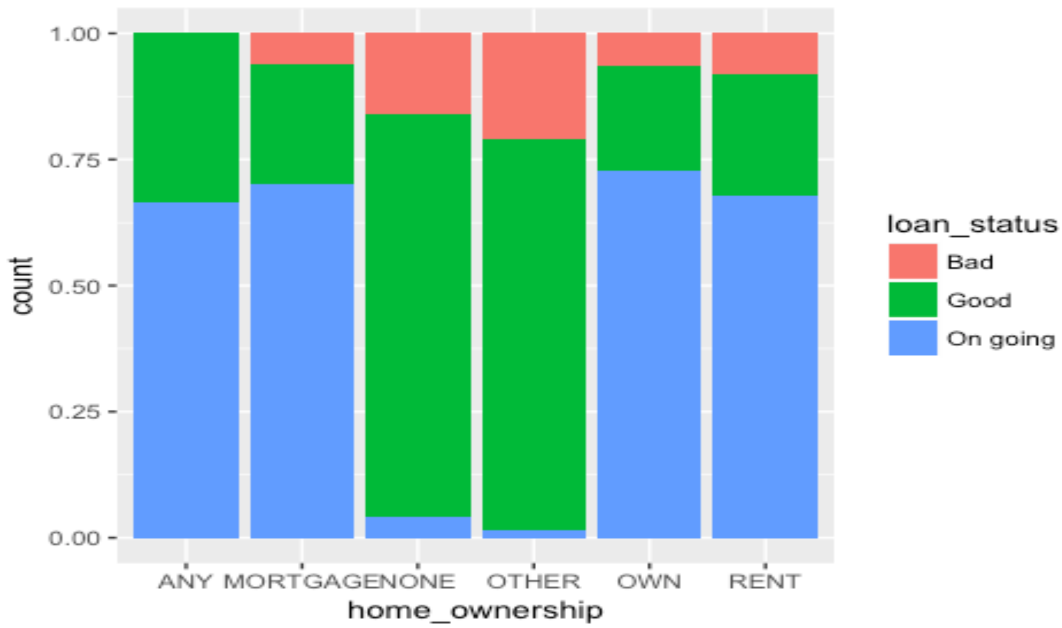
individual has much more proportion of bad status,may be because many joint loans are not finished.

```
#term
ggplot(X,aes(term,fill=loan_status))+geom_bar(position = "fill")
```



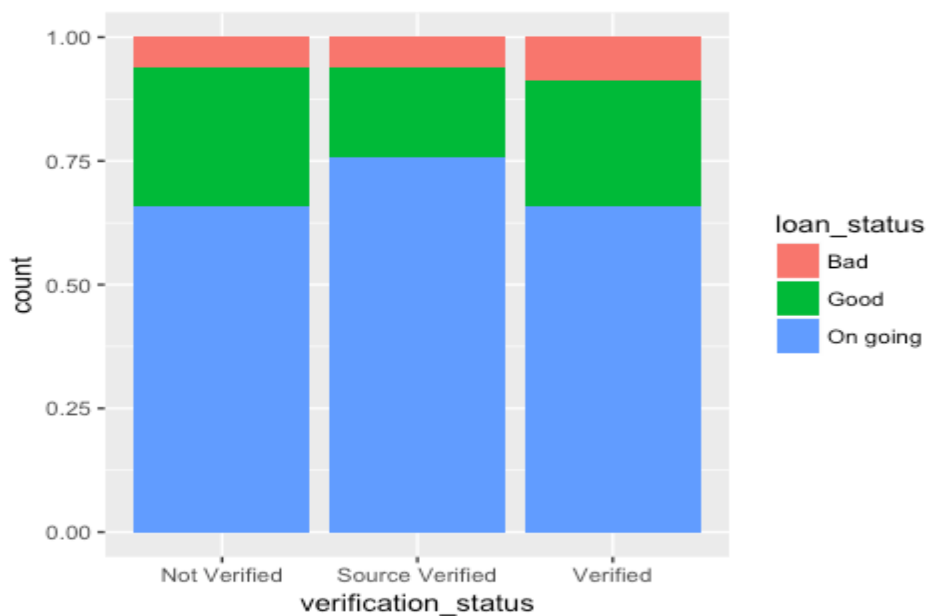
long term loans has more proportion of bad status.

```r
#home_ownership
ggplot(X,aes(home_ownership,fill=loan_status))+geom_bar(position =
"fill")
```



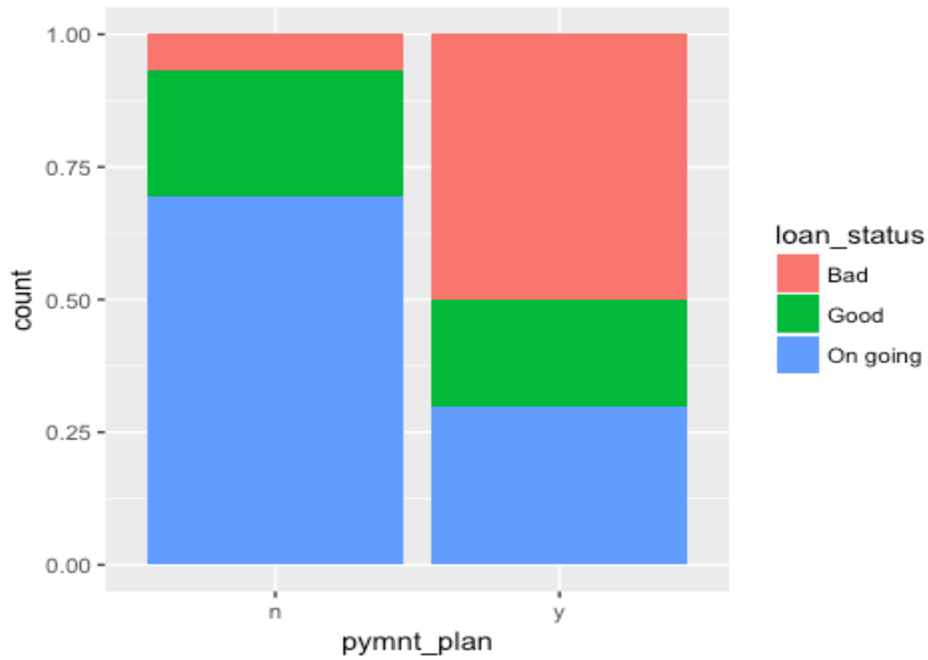
none or "others" has more proportion of bad status, but also more good loans

```r
#verification
ggplot(X,aes(verification_status,fill=loan_status))+geom_bar(position =
"fill")
```



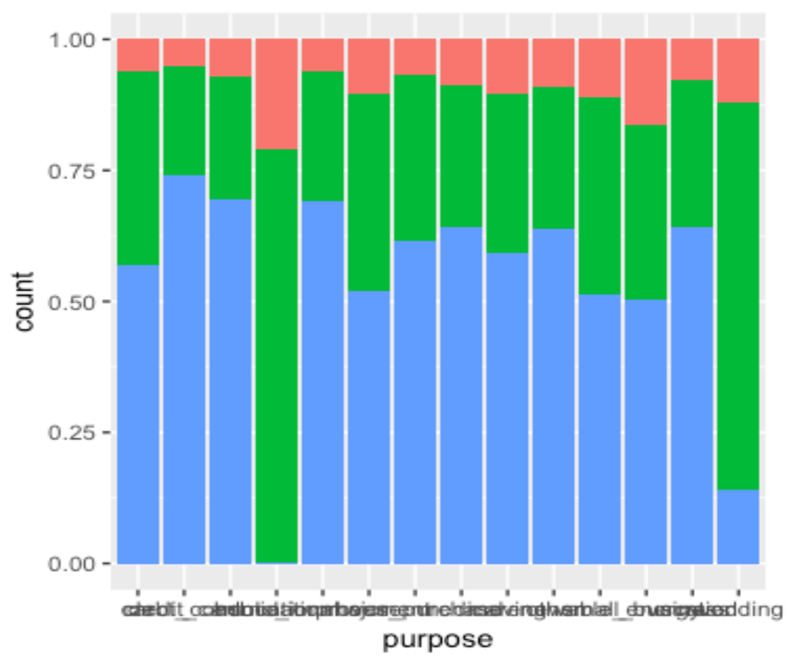
those who are verified have more proportion of bad status instead. It may suggest that this variable is not an good indicator of loan_status.

```r
#payment plan
ggplot(X,aes(pymnt_plan,fill=loan_status))+geom_bar(position = "fill")
```



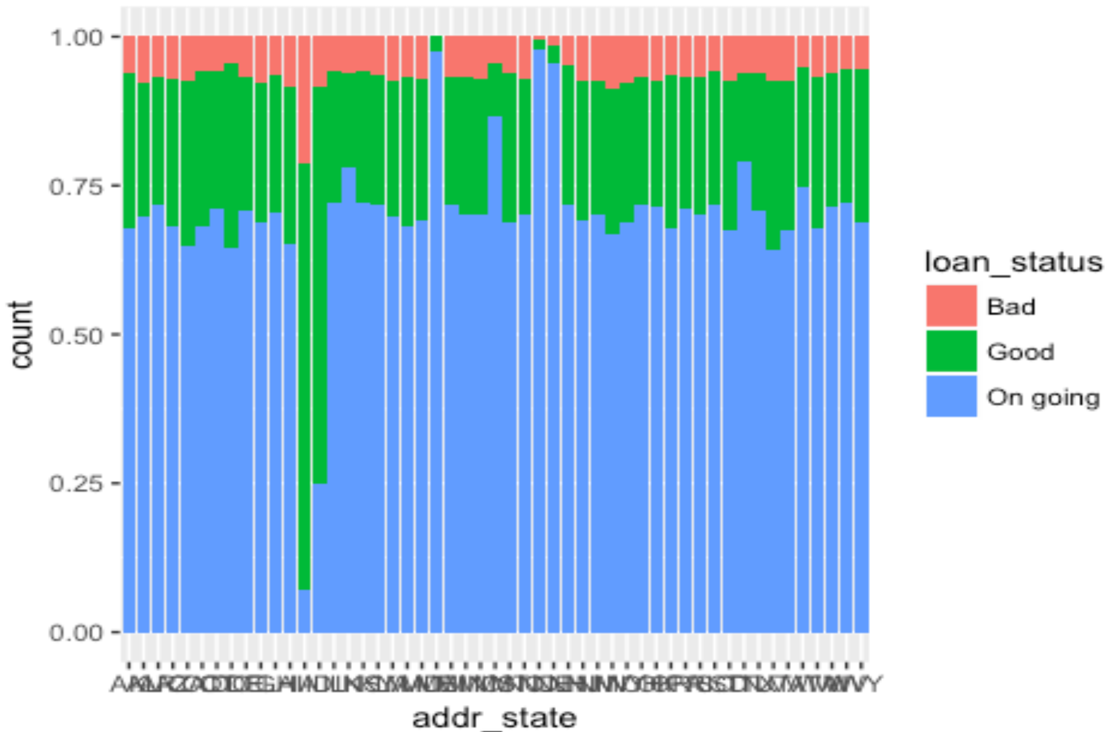
half of those having plan are actually in bad status. However, the total count is 10 which is too small to consider.

```r
#purpose
ggplot(X,aes(purpose,fill=loan_status))+geom_bar(position = "fill")
```



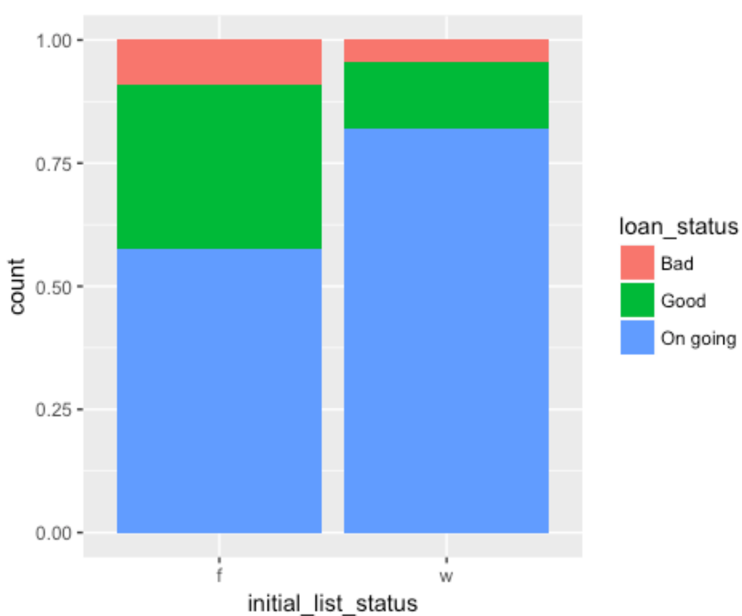
educational, small_business have much higher rates of bad status than other purposes.

```r
ggplot(X,aes(addr_state,fill=loan_status))+geom_bar(position = "fill")
```



ME,ND,NE has low rates of bad status. IA has high rates but there are only 14 counts in total. This variable seems significant.

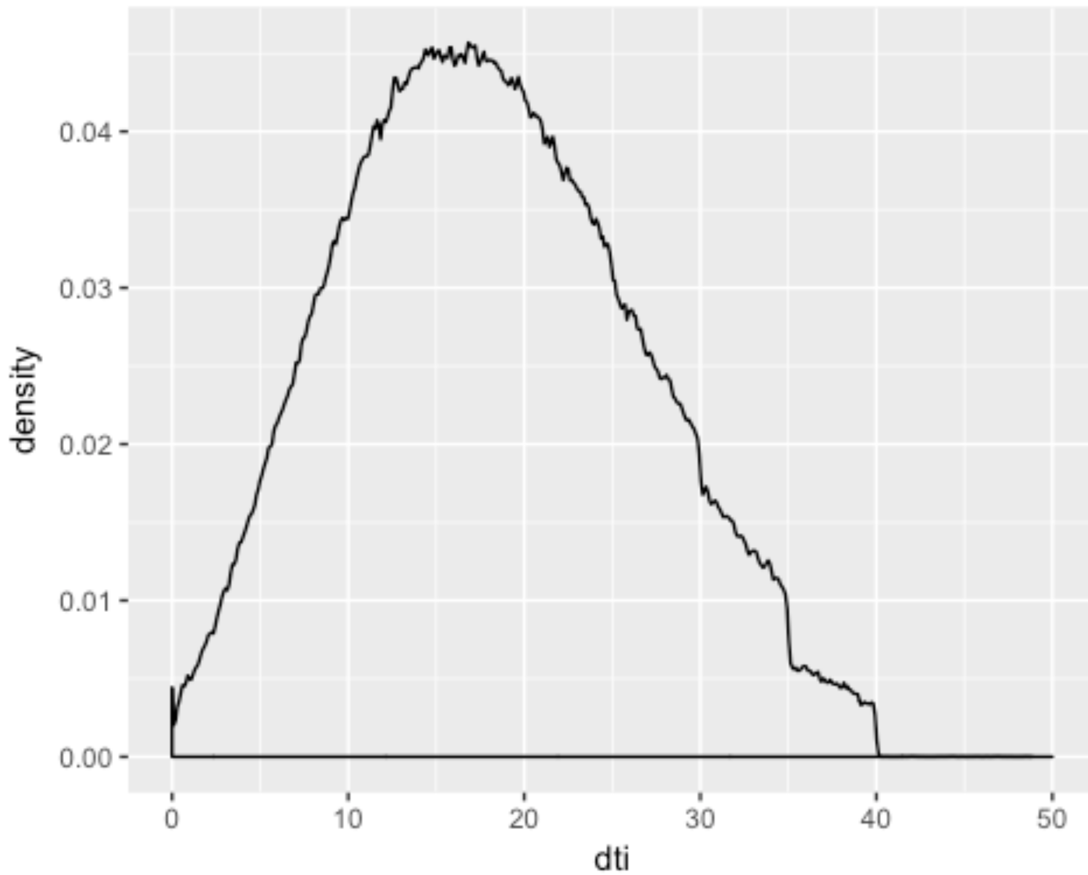
```r
ggplot(X,aes(initial_list_status,fill=loan_status))+geom_bar(position =
"fill")
```



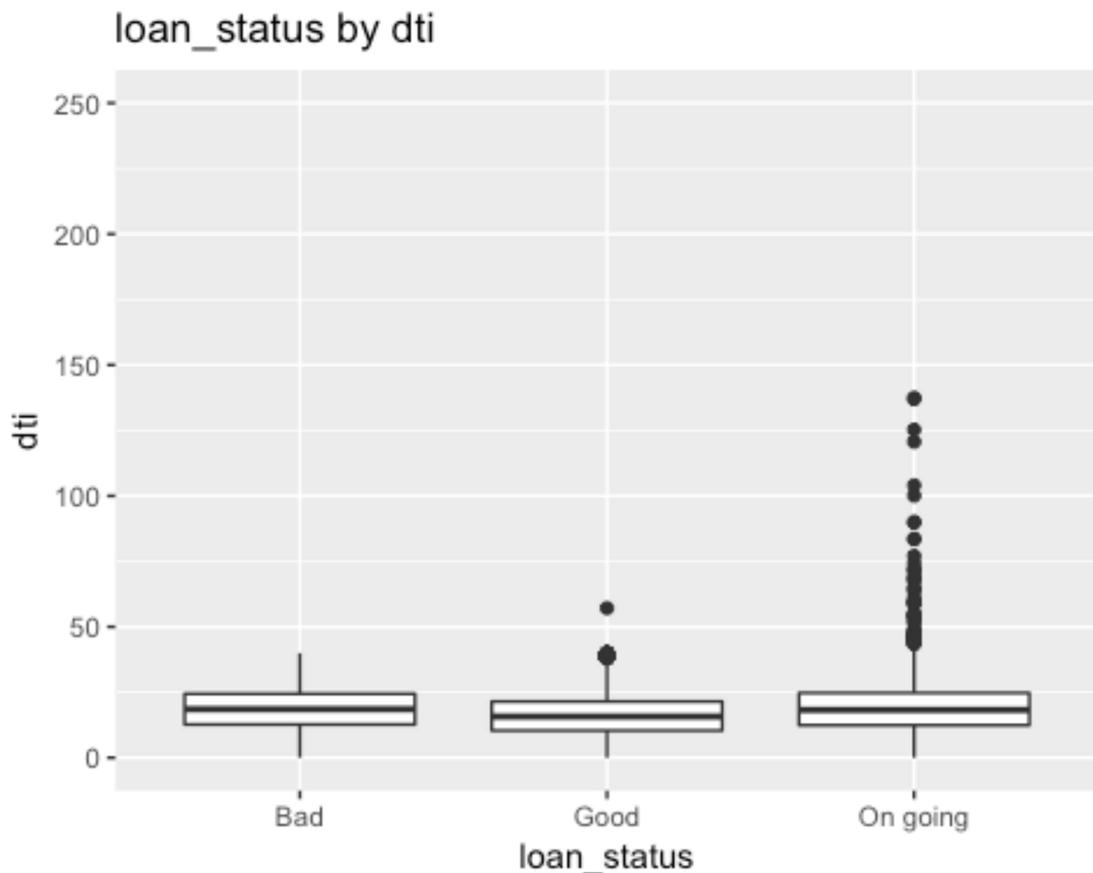
f has high rates of bad status. but w also have more ongoing loans.

continuous variables and loan_status

```
#dti seems to be an important data based on explanation from dictionary
ggplot(X, aes(dti))+geom_density(bw=0.05)+xlim(c(0,50))

## Warning: Removed 44 rows containing non-finite values
(stat_density).
```




```
box_plane = ggplot(X, aes(loan_status,dti))+ylim(c(0,250))
box_plane + geom_boxplot(aes(fill = dti)) +
  labs(title = "loan_status by dti",
       x = "loan_status",
       y = "dti")

## Warning: Removed 5 rows containing non-finite values (stat_boxplot).
```



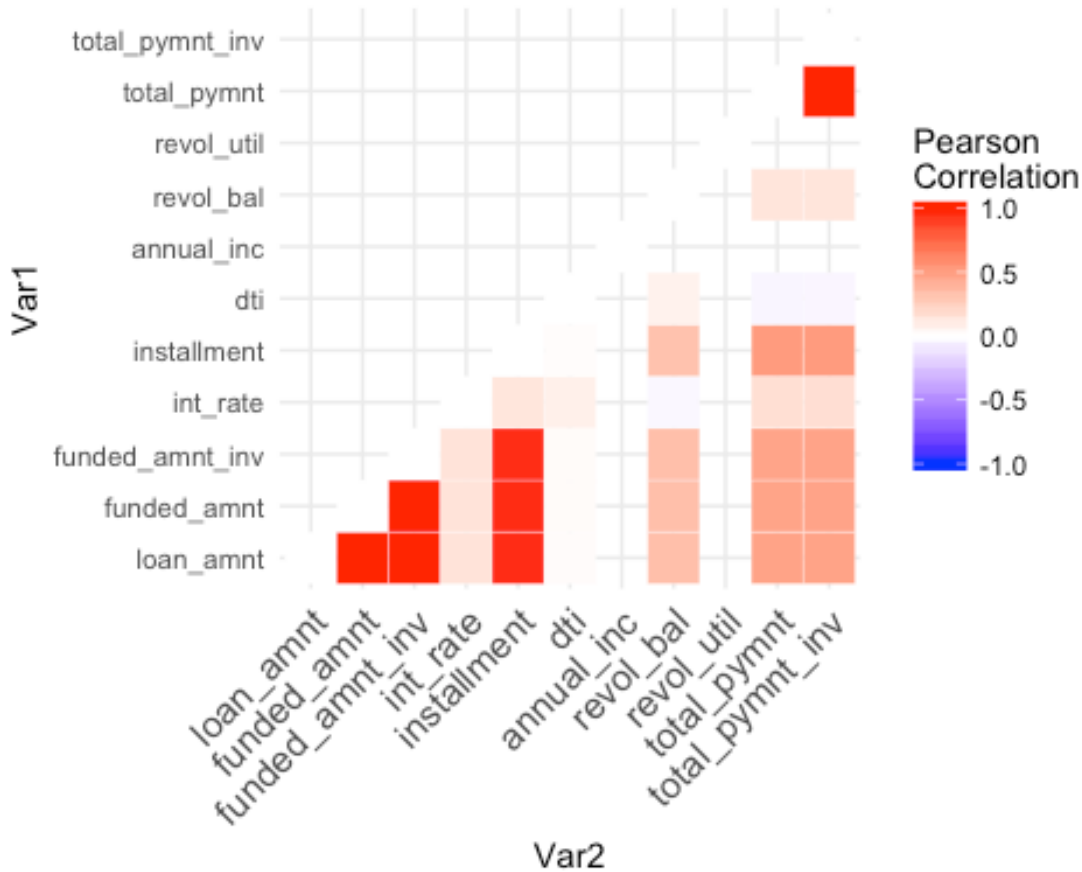
the result shows that dti is lower for good status, which corresponds to our prediction, indicates that dti is an important feature.

explore correlation among continuous variables

```
#construct the correlation matrix for some variables
cormat = cor (X[, c("loan_amnt","funded_amnt", "funded_amnt_inv",
"int_rate","installment", "dti", "annual_inc","revol_bal",
                    "revol_util", "total_pymnt","total_pymnt_inv")])

#Remove self correlations
diag (cormat) = 0
cormat

##                  loan_amnt funded_amnt funded_amnt_inv    int_rate
## loan_amnt       0.00000000  0.99926263      0.99711526  0.14502310
## funded_amnt     0.99926263  0.00000000      0.99802509  0.14516034
## funded_amnt_inv 0.99711526  0.99802509      0.00000000  0.14520528
## int_rate        0.14502310  0.14516034      0.14520528  0.00000000
## installment     0.94497724  0.94600491      0.94363202  0.13307492
## dti             0.02067549  0.02107492      0.02218536  0.07990255
## annual_inc              NA          NA              NA          NA
## revol_bal       0.33357999  0.33343530      0.33173609 -0.03570809
## revol_util              NA          NA              NA          NA
```

```
## total_pymnt     0.47462594  0.47328577      0.46884829  0.17050629
## total_pymnt_inv 0.47565520  0.47450204      0.47406155  0.17147933
##                 installment         dti annual_inc   revol_bal
revol_util
## loan_amnt        0.94497724  0.02067549         NA  0.33357999
NA
## funded_amnt      0.94600491  0.02107492         NA  0.33343530
NA
## funded_amnt_inv  0.94363202  0.02218536         NA  0.33173609
NA
## int_rate         0.13307492  0.07990255         NA -0.03570809
NA
## installment      0.00000000  0.01433284         NA  0.31658819
NA
## dti              0.01433284  0.00000000         NA  0.06727728
NA
## annual_inc               NA          NA          0          NA
NA
## revol_bal        0.31658819  0.06727728         NA  0.00000000
NA
## revol_util               NA          NA         NA          NA
0
## total_pymnt      0.51495367 -0.04152877         NA  0.13832761
NA
## total_pymnt_inv  0.51581715 -0.04033598         NA  0.13774610
NA
##                 total_pymnt total_pymnt_inv
## loan_amnt        0.47462594      0.47565520
## funded_amnt      0.47328577      0.47450204
## funded_amnt_inv  0.46884829      0.47406155
## int_rate         0.17050629      0.17147933
## installment      0.51495367      0.51581715
## dti             -0.04152877     -0.04033598
## annual_inc               NA              NA
## revol_bal        0.13832761      0.13774610
## revol_util               NA              NA
## total_pymnt      0.00000000      0.99759232
## total_pymnt_inv  0.99759232      0.00000000

get_upper_tri <- function(cormat){
  cormat[lower.tri(cormat)]<- NA
  return(cormat)
}
upper_tri <- get_upper_tri(cormat)

library(reshape2)
melted_cormat <- melt(upper_tri, na.rm = TRUE)

# plot correlation heatmap
ggplot(data = melted_cormat, aes(Var2, Var1, fill = value))+
```

```
  geom_tile(color = "white")+
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                       midpoint = 0, limit = c(-1,1), space = "Lab",
                       name="Pearson\nCorrelation") +
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 45, vjust = 1,
                                   size = 12, hjust = 1))+
  coord_fixed()
```



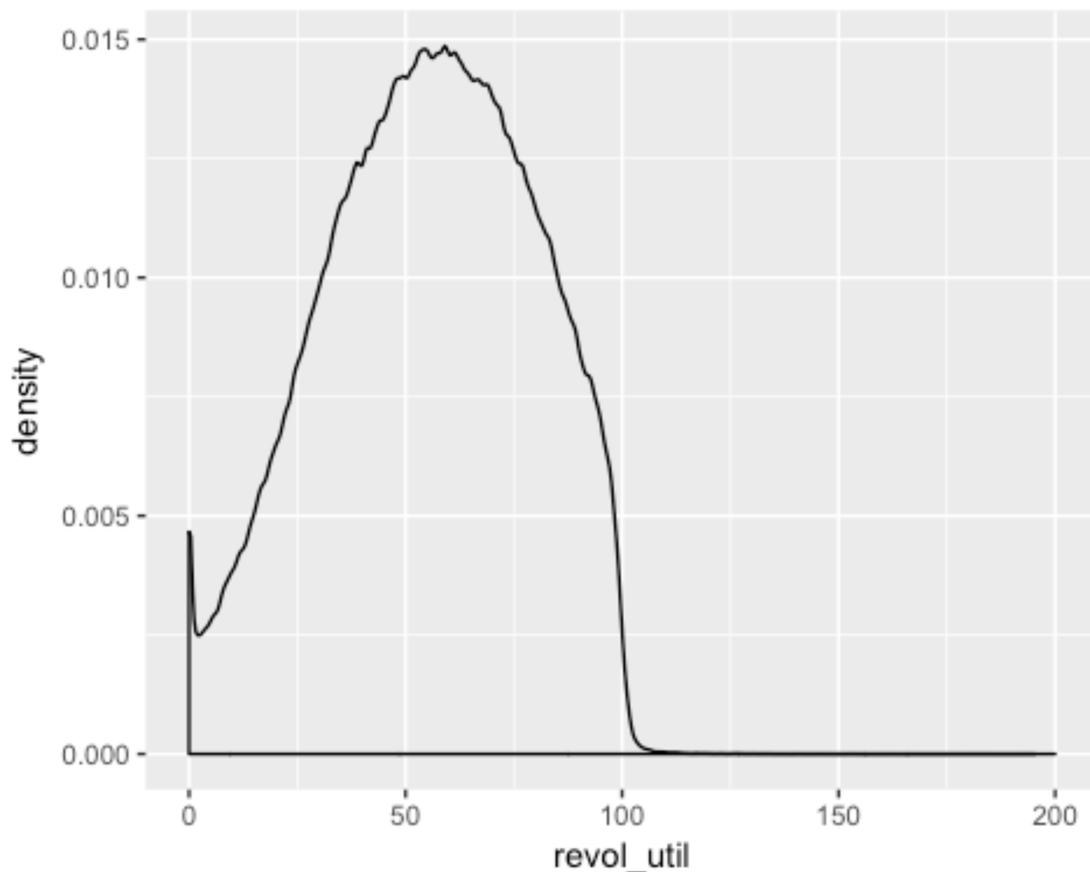
Strangely, annual_inc and revol_util have no correlation with other variables. loan_amnt, funded_amnt and funded_amnt_inv have high correlation as expected, it will be reasonable to choose two out of three or one out of three.

4. Dealing with missing data

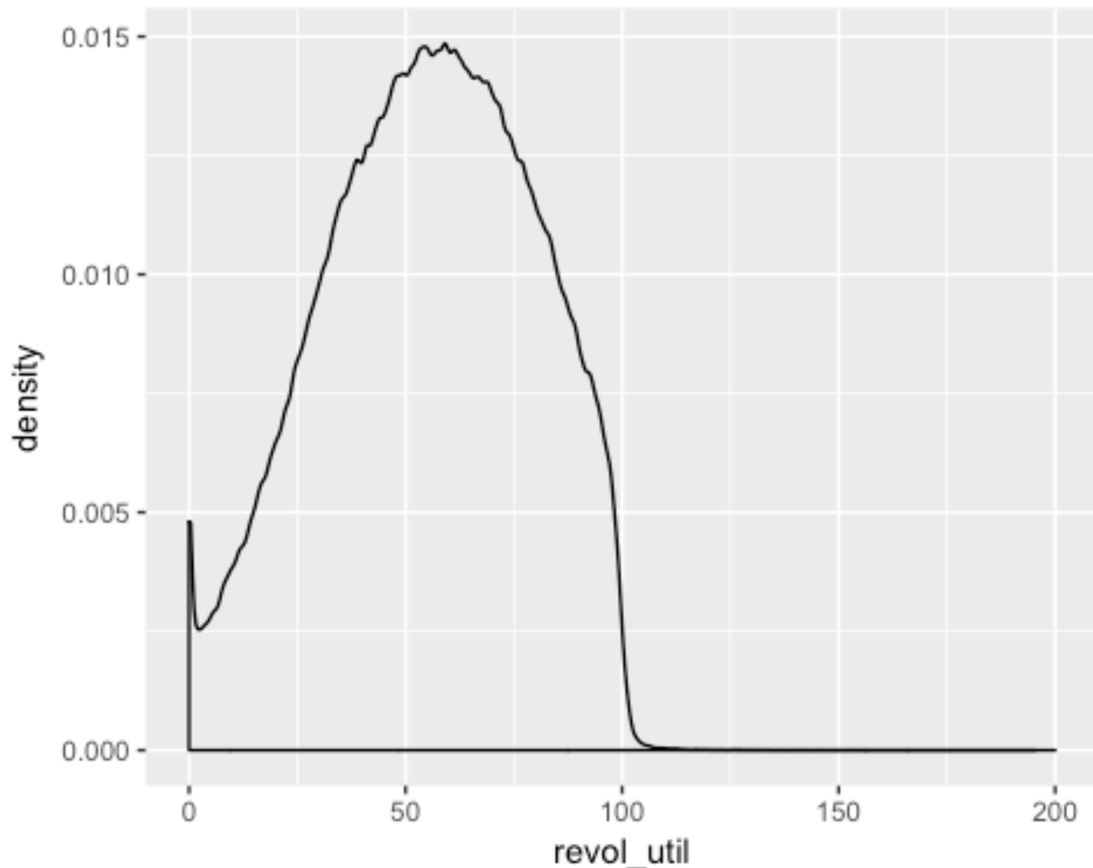
There are many missing data in this dataset. I use summary() to explore those variables, and try to impute one variable using the Mice package.

```
#incomplete data related to behaviours of the borrower
sum(is.na(X$revol_util))

## [1] 502

#502 NAs
#impute the number of revol_util
# Set a random seed
set.seed(129)

# Perform mice imputation, excluding certain less-than-useful
variables:
library(mice)

## Warning: package 'mice' was built under R version 3.3.2

mice_mod <- mice(X[, names(X) %in% c("revol_bal","revol_util")],
method='pmm')

##
##  iter imp variable
##   1   1  revol_util
##   1   2  revol_util
##   1   3  revol_util
##   1   4  revol_util
##   1   5  revol_util
##   2   1  revol_util
##   2   2  revol_util
##   2   3  revol_util
##   2   4  revol_util
##   2   5  revol_util
##   3   1  revol_util
##   3   2  revol_util
##   3   3  revol_util
##   3   4  revol_util
##   3   5  revol_util
##   4   1  revol_util
##   4   2  revol_util
##   4   3  revol_util
##   4   4  revol_util
##   4   5  revol_util
##   5   1  revol_util
##   5   2  revol_util
##   5   3  revol_util
##   5   4  revol_util
##   5   5  revol_util
```

```r
#at first I tried the whole dataset and method=rf, but it takes too
long, so I use pmm instead
mice_output <- complete(mice_mod)
#compare the output with original age data
# Plot revol_util distributions
par(mfrow=c(1,2))
ggplot(X, aes(revol_util))+geom_density(bw=0.5)+xlim(0,200)

## Warning: Removed 504 rows containing non-finite values
(stat_density).
```




```r
ggplot(mice_output, aes(revol_util))+geom_density(bw=0.5)+xlim(0,200)

## Warning: Removed 2 rows containing non-finite values (stat_density).
```



Since the two graphs are similar, it is safe to use the imputed data to fill the NAs

```
X$revol_util <- mice_output$revol_util
# Show new number of missing values
sum(is.na(X$revol_util))

## [1] 0
```

However, I realise that imputation takes a long time for this big dataset and it will not contribute too much for the model if I sacrifice running time and use only one variable to predict the NAs in another variable. Besides, by examining the summary of the other variables, I noticed that many variables have almost half of their data missing, so it may be too risky to impute values for these variables.

5. Decisions to exclude some variables from the final model

```
#create a new dataframe for model testing
#variables that are not important: member_id,emp_title, url
Y<-X[,!names(X)%in%c("member_id","emp_title", "url")]
```

variables needed further research, but not in this discussion(more discussion in evaluation):

desc(natural language analytics), title(correlated with purpose)

```
Y<-Y[,!names(Y)%in%c("desc","title")]
```

variables that may introduce confounding effect: "out_prncp","out_prncp_inv","total_pymnt","total_pymnt_inv","total_rec_prncp","total_rec_int","recoveries","collection_recovery_fee","last_pymnt_amnt","last_pymnt_d","next_pymnt_d"

These variables will affect model building process and they does not contribute much to our understanding of the study as a whole. For example, if "recoveries">0, it means that the status is likely to be charged off. It does not help with our analysis because we certainly know that recoveries will only exist if payment was not made in time. Therefore, I discard these variables from the model.

```
Y<-
Y[,!names(Y)%in%c("issue_d","out_prncp","out_prncp_inv","total_pymnt","total_pymnt_inv"
                     ,"total_rec_prncp","total_rec_int","recoveries",

"collection_recovery_fee","last_pymnt_amnt","last_pymnt_d","next_pymnt_d")]
```

Remove "joint" appliation type, as well as variables related to joint application: "annual_inc_joint","dti_joint","verification_status_joint"

Rationale: I found out that there are only 511 cases of "joint" application type, which is a very small sample as compared to "individual". Instead of creating a model for both types, I feel that it gives more accurate result to construct a model for "individual" since there are enough samples. As for "joint", maybe we can try to collect more sample or choose less cases from "individual". For this study, I will only focus on "individual".

```
Y<-Y[Y$application_type=="INDIVIDUAL",]
Y<-
Y[,!names(Y)%in%c("annual_inc_joint","dti_joint","verification_status_joint")]
```

Drop variables which are subsituted by new variables:

earliest_cr_line, last_credit_pull_d

```
Y<-Y[,!names(Y)%in%c("earliest_cr_line", "last_credit_pull_d")]
```

variables which are highly correlated: "funded_amnt"

```
Y<-Y[,!names(Y)%in%("funded_amnt")]
```

factorise some variables:

```
factor_vars <-

c("delinq_2yrs","inq_last_6mths","mths_since_last_delinq","mths_since_l
ast_record",

"open_acc","pub_rec","total_acc","mths_since_last_major_derog","policy_
code","acc_now_delinq",
    "open_acc_6m","open_il_6m", "open_il_12m","open_il_24m",
"mths_since_rcnt_il",
    "open_rv_12m","open_rv_24m"
,"inq_fi","inq_last_12m","months_from_earliest_cr_line",
    "months_from_last_credit_pull_d")

Y[factor_vars] <- lapply(Y[factor_vars], function(x) as.factor(x))
```

6. model building

```
#separate into training set and testing set
n_total = length(Y[,1])
trainindex= sample(1:n_total, 10000)
testindex= sample(1:n_total, 10000)
Ytrain<-Y[trainindex,]
Ynotrain<-Y[-trainindex,]
Ytest<-Ynotrain[testindex,]
```

At first I separate the training set and the testing set equally based on the whole dataset, but afterwards I realised that my laptop simply cannot finish the computation with this many data. So I take a small sample for the purpose of this analysis.

I chose Xgboosting because it computes faster and gives good result.

```
library(xgboost)

## Warning: package 'xgboost' was built under R version 3.3.2

library(readr)
library(stringr)

## Warning: package 'stringr' was built under R version 3.3.2

library(caret)

## Warning: package 'caret' was built under R version 3.3.2
```

```
## Loading required package: lattice

library(car)

## Warning: package 'car' was built under R version 3.3.2

##
## Attaching package: 'car'

## The following object is masked from 'package:DescTools':
##
##     Recode

xgb <- xgboost(data = data.matrix(Ytrain[,-c(1,13)]), #without ID and
loan_status
               label = as.numeric(Ytrain$loan_status)-1,
               eta = 0.01,
               max_depth = 15,
               nround=1000,
               subsample = 0.5,
               colsample_bytree = 0.5,
               seed = 1,
               eval_metric = "merror",
               objective = "multi:softmax",
               num_class = 3,
               nthread = 3
)

## [1]  train-merror:0.209900
## [2]  train-merror:0.171700
## [3]  train-merror:0.164200
## [4]  train-merror:0.163000
## [5]  train-merror:0.154100
## [6]  train-merror:0.149600
## [7]  train-merror:0.149200
## [8]  train-merror:0.148600
## [9]  train-merror:0.150900
## [10] train-merror:0.146100
## [11] train-merror:0.146000
## [12] train-merror:0.145100
## [13] train-merror:0.142600
## [14] train-merror:0.142700
## [15] train-merror:0.143000
## [16] train-merror:0.140100
## [17] train-merror:0.140000
## [18] train-merror:0.139300
## [19] train-merror:0.137900
## [20] train-merror:0.137300
## [21] train-merror:0.136500
## [22] train-merror:0.136500
## [23] train-merror:0.135900
```

```
## [24] train-merror:0.136100
## [25] train-merror:0.135600
## [26] train-merror:0.135400
## [27] train-merror:0.136500
## [28] train-merror:0.136100
## [29] train-merror:0.135700
## [30] train-merror:0.134700
## [31] train-merror:0.133500
## [32] train-merror:0.133000
## [33] train-merror:0.131900
## [34] train-merror:0.132200
## [35] train-merror:0.132000
## [36] train-merror:0.131700
## [37] train-merror:0.131000
## [38] train-merror:0.130000
## [39] train-merror:0.129900
## [40] train-merror:0.130200
## [41] train-merror:0.129700
## [42] train-merror:0.128500
## [43] train-merror:0.127200
## [44] train-merror:0.128000
## [45] train-merror:0.128000
## [46] train-merror:0.128600
## [47] train-merror:0.127400
## [48] train-merror:0.127000
## [49] train-merror:0.125600
## [50] train-merror:0.125500
## [51] train-merror:0.125300
## [52] train-merror:0.124800
## [53] train-merror:0.123600
## [54] train-merror:0.123800
## [55] train-merror:0.122900
## [56] train-merror:0.122400
## [57] train-merror:0.122400
## [58] train-merror:0.121300
## [59] train-merror:0.121700
## [60] train-merror:0.121000
## [61] train-merror:0.119900
## [62] train-merror:0.119900
## [63] train-merror:0.119500
## [64] train-merror:0.119500
## [65] train-merror:0.118800
## [66] train-merror:0.118600
## [67] train-merror:0.118200
## [68] train-merror:0.117500
## [69] train-merror:0.116800
## [70] train-merror:0.115500
## [71] train-merror:0.115500
## [72] train-merror:0.114700
## [73] train-merror:0.114300
```

```
## [74] train-merror:0.115400
## [75] train-merror:0.114500
## [76] train-merror:0.114500
## [77] train-merror:0.113600
## [78] train-merror:0.113100
## [79] train-merror:0.113200
## [80] train-merror:0.112500
## [81] train-merror:0.112200
## [82] train-merror:0.112200
## [83] train-merror:0.111900
## [84] train-merror:0.111800
## [85] train-merror:0.110900
## [86] train-merror:0.110200
## [87] train-merror:0.109800
## [88] train-merror:0.109400
## [89] train-merror:0.108700
## [90] train-merror:0.108000
## [91] train-merror:0.107500
## [92] train-merror:0.106500
## [93] train-merror:0.106100
## [94] train-merror:0.106300
## [95] train-merror:0.105800
## [96] train-merror:0.105900
## [97] train-merror:0.104700
## [98] train-merror:0.103900
## [99] train-merror:0.103500
## [100]    train-merror:0.103100
## [101]    train-merror:0.103400
## [102]    train-merror:0.103500
## [103]    train-merror:0.102400
## [104]    train-merror:0.102500
## [105]    train-merror:0.102200
## [106]    train-merror:0.101800
## [107]    train-merror:0.101600
## [108]    train-merror:0.101000
## [109]    train-merror:0.100900
## [110]    train-merror:0.100300
## [111]    train-merror:0.099700
## [112]    train-merror:0.099300
## [113]    train-merror:0.099700
## [114]    train-merror:0.099100
## [115]    train-merror:0.099000
## [116]    train-merror:0.098700
## [117]    train-merror:0.098000
## [118]    train-merror:0.097900
## [119]    train-merror:0.097500
## [120]    train-merror:0.097200
## [121]    train-merror:0.096400
## [122]    train-merror:0.096400
## [123]    train-merror:0.096000
```

```
## [124]	train-merror:0.095400
## [125]	train-merror:0.095100
## [126]	train-merror:0.095000
## [127]	train-merror:0.094700
## [128]	train-merror:0.093800
## [129]	train-merror:0.093600
## [130]	train-merror:0.093800
## [131]	train-merror:0.093200
## [132]	train-merror:0.093500
## [133]	train-merror:0.092900
## [134]	train-merror:0.092800
## [135]	train-merror:0.092000
## [136]	train-merror:0.091200
## [137]	train-merror:0.090900
## [138]	train-merror:0.090700
## [139]	train-merror:0.089800
## [140]	train-merror:0.089300
## [141]	train-merror:0.089100
## [142]	train-merror:0.089200
## [143]	train-merror:0.089000
## [144]	train-merror:0.088500
## [145]	train-merror:0.088500
## [146]	train-merror:0.088000
## [147]	train-merror:0.087400
## [148]	train-merror:0.087200
## [149]	train-merror:0.086900
## [150]	train-merror:0.086700
## [151]	train-merror:0.086700
## [152]	train-merror:0.086700
## [153]	train-merror:0.086100
## [154]	train-merror:0.085900
## [155]	train-merror:0.085000
## [156]	train-merror:0.085100
## [157]	train-merror:0.084900
## [158]	train-merror:0.084200
## [159]	train-merror:0.084400
## [160]	train-merror:0.083800
## [161]	train-merror:0.083500
## [162]	train-merror:0.083000
## [163]	train-merror:0.083000
## [164]	train-merror:0.082200
## [165]	train-merror:0.081600
## [166]	train-merror:0.081800
## [167]	train-merror:0.081100
## [168]	train-merror:0.081300
## [169]	train-merror:0.080600
## [170]	train-merror:0.080700
## [171]	train-merror:0.080000
## [172]	train-merror:0.078900
## [173]	train-merror:0.078800
```

```
## [174]	train-merror:0.078500
## [175]	train-merror:0.078700
## [176]	train-merror:0.078200
## [177]	train-merror:0.077800
## [178]	train-merror:0.077500
## [179]	train-merror:0.077300
## [180]	train-merror:0.077400
## [181]	train-merror:0.077000
## [182]	train-merror:0.076900
## [183]	train-merror:0.076400
## [184]	train-merror:0.076000
## [185]	train-merror:0.075700
## [186]	train-merror:0.075200
## [187]	train-merror:0.074400
## [188]	train-merror:0.074600
## [189]	train-merror:0.073800
## [190]	train-merror:0.073200
## [191]	train-merror:0.073300
## [192]	train-merror:0.072800
## [193]	train-merror:0.072400
## [194]	train-merror:0.072700
## [195]	train-merror:0.072200
## [196]	train-merror:0.072300
## [197]	train-merror:0.072100
## [198]	train-merror:0.071500
## [199]	train-merror:0.071100
## [200]	train-merror:0.070700
## [201]	train-merror:0.070100
## [202]	train-merror:0.069900
## [203]	train-merror:0.069700
## [204]	train-merror:0.069500
## [205]	train-merror:0.069200
## [206]	train-merror:0.069300
## [207]	train-merror:0.068500
## [208]	train-merror:0.068400
## [209]	train-merror:0.068200
## [210]	train-merror:0.068300
## [211]	train-merror:0.067800
## [212]	train-merror:0.067500
## [213]	train-merror:0.066800
## [214]	train-merror:0.066500
## [215]	train-merror:0.065800
## [216]	train-merror:0.065900
## [217]	train-merror:0.065500
## [218]	train-merror:0.065400
## [219]	train-merror:0.064700
## [220]	train-merror:0.064600
## [221]	train-merror:0.064400
## [222]	train-merror:0.064100
## [223]	train-merror:0.063800
```

```
## [224]	train-merror:0.063600
## [225]	train-merror:0.063600
## [226]	train-merror:0.063500
## [227]	train-merror:0.063000
## [228]	train-merror:0.062600
## [229]	train-merror:0.062200
## [230]	train-merror:0.061800
## [231]	train-merror:0.061600
## [232]	train-merror:0.061500
## [233]	train-merror:0.061000
## [234]	train-merror:0.061000
## [235]	train-merror:0.060400
## [236]	train-merror:0.060600
## [237]	train-merror:0.060200
## [238]	train-merror:0.059700
## [239]	train-merror:0.059400
## [240]	train-merror:0.059100
## [241]	train-merror:0.058800
## [242]	train-merror:0.058500
## [243]	train-merror:0.058300
## [244]	train-merror:0.058200
## [245]	train-merror:0.057600
## [246]	train-merror:0.057600
## [247]	train-merror:0.057200
## [248]	train-merror:0.057200
## [249]	train-merror:0.057200
## [250]	train-merror:0.057200
## [251]	train-merror:0.057100
## [252]	train-merror:0.056400
## [253]	train-merror:0.056300
## [254]	train-merror:0.055900
## [255]	train-merror:0.055600
## [256]	train-merror:0.055500
## [257]	train-merror:0.055600
## [258]	train-merror:0.055300
## [259]	train-merror:0.055000
## [260]	train-merror:0.054600
## [261]	train-merror:0.054400
## [262]	train-merror:0.054500
## [263]	train-merror:0.053800
## [264]	train-merror:0.053400
## [265]	train-merror:0.053400
## [266]	train-merror:0.053400
## [267]	train-merror:0.053000
## [268]	train-merror:0.052700
## [269]	train-merror:0.052300
## [270]	train-merror:0.051900
## [271]	train-merror:0.051500
## [272]	train-merror:0.051700
## [273]	train-merror:0.051100
```

```
## [274]	train-merror:0.050500
## [275]	train-merror:0.050300
## [276]	train-merror:0.049900
## [277]	train-merror:0.050100
## [278]	train-merror:0.049400
## [279]	train-merror:0.050000
## [280]	train-merror:0.049500
## [281]	train-merror:0.049200
## [282]	train-merror:0.049200
## [283]	train-merror:0.048800
## [284]	train-merror:0.048600
## [285]	train-merror:0.048200
## [286]	train-merror:0.048200
## [287]	train-merror:0.047400
## [288]	train-merror:0.047400
## [289]	train-merror:0.047200
## [290]	train-merror:0.047000
## [291]	train-merror:0.046800
## [292]	train-merror:0.046800
## [293]	train-merror:0.046500
## [294]	train-merror:0.046200
## [295]	train-merror:0.045500
## [296]	train-merror:0.045100
## [297]	train-merror:0.045100
## [298]	train-merror:0.045000
## [299]	train-merror:0.044600
## [300]	train-merror:0.044800
## [301]	train-merror:0.044600
## [302]	train-merror:0.044100
## [303]	train-merror:0.043900
## [304]	train-merror:0.043500
## [305]	train-merror:0.043300
## [306]	train-merror:0.043100
## [307]	train-merror:0.043100
## [308]	train-merror:0.042500
## [309]	train-merror:0.042300
## [310]	train-merror:0.042100
## [311]	train-merror:0.042300
## [312]	train-merror:0.041900
## [313]	train-merror:0.041400
## [314]	train-merror:0.041300
## [315]	train-merror:0.041000
## [316]	train-merror:0.040600
## [317]	train-merror:0.040700
## [318]	train-merror:0.040300
## [319]	train-merror:0.040300
## [320]	train-merror:0.040200
## [321]	train-merror:0.039600
## [322]	train-merror:0.039500
## [323]	train-merror:0.039000
```

```
## [324]	train-merror:0.039300
## [325]	train-merror:0.039500
## [326]	train-merror:0.039000
## [327]	train-merror:0.038400
## [328]	train-merror:0.038200
## [329]	train-merror:0.038300
## [330]	train-merror:0.037800
## [331]	train-merror:0.037900
## [332]	train-merror:0.037300
## [333]	train-merror:0.037300
## [334]	train-merror:0.036900
## [335]	train-merror:0.036400
## [336]	train-merror:0.036600
## [337]	train-merror:0.036200
## [338]	train-merror:0.035800
## [339]	train-merror:0.035700
## [340]	train-merror:0.035500
## [341]	train-merror:0.035300
## [342]	train-merror:0.035100
## [343]	train-merror:0.035000
## [344]	train-merror:0.034400
## [345]	train-merror:0.034500
## [346]	train-merror:0.034500
## [347]	train-merror:0.034400
## [348]	train-merror:0.034000
## [349]	train-merror:0.034000
## [350]	train-merror:0.033700
## [351]	train-merror:0.033400
## [352]	train-merror:0.033400
## [353]	train-merror:0.032800
## [354]	train-merror:0.032800
## [355]	train-merror:0.032400
## [356]	train-merror:0.032600
## [357]	train-merror:0.032500
## [358]	train-merror:0.032200
## [359]	train-merror:0.032000
## [360]	train-merror:0.031800
## [361]	train-merror:0.031100
## [362]	train-merror:0.031500
## [363]	train-merror:0.031300
## [364]	train-merror:0.031300
## [365]	train-merror:0.030900
## [366]	train-merror:0.030600
## [367]	train-merror:0.030700
## [368]	train-merror:0.030400
## [369]	train-merror:0.030000
## [370]	train-merror:0.029900
## [371]	train-merror:0.029800
## [372]	train-merror:0.029700
## [373]	train-merror:0.029400
```

```
## [374]	train-merror:0.029200
## [375]	train-merror:0.029200
## [376]	train-merror:0.029000
## [377]	train-merror:0.028900
## [378]	train-merror:0.028700
## [379]	train-merror:0.028300
## [380]	train-merror:0.027900
## [381]	train-merror:0.027400
## [382]	train-merror:0.027600
## [383]	train-merror:0.027400
## [384]	train-merror:0.027600
## [385]	train-merror:0.027400
## [386]	train-merror:0.027100
## [387]	train-merror:0.026700
## [388]	train-merror:0.026600
## [389]	train-merror:0.026400
## [390]	train-merror:0.026300
## [391]	train-merror:0.026000
## [392]	train-merror:0.026100
## [393]	train-merror:0.025600
## [394]	train-merror:0.024900
## [395]	train-merror:0.024700
## [396]	train-merror:0.024700
## [397]	train-merror:0.024600
## [398]	train-merror:0.024300
## [399]	train-merror:0.024300
## [400]	train-merror:0.024100
## [401]	train-merror:0.024100
## [402]	train-merror:0.023900
## [403]	train-merror:0.024000
## [404]	train-merror:0.023900
## [405]	train-merror:0.023900
## [406]	train-merror:0.023700
## [407]	train-merror:0.023600
## [408]	train-merror:0.023700
## [409]	train-merror:0.023300
## [410]	train-merror:0.023400
## [411]	train-merror:0.023000
## [412]	train-merror:0.022800
## [413]	train-merror:0.022700
## [414]	train-merror:0.022600
## [415]	train-merror:0.022300
## [416]	train-merror:0.022200
## [417]	train-merror:0.022100
## [418]	train-merror:0.022000
## [419]	train-merror:0.021600
## [420]	train-merror:0.021400
## [421]	train-merror:0.021400
## [422]	train-merror:0.021300
## [423]	train-merror:0.021200
```

```
## [424]	train-merror:0.020900
## [425]	train-merror:0.021000
## [426]	train-merror:0.020700
## [427]	train-merror:0.020700
## [428]	train-merror:0.020500
## [429]	train-merror:0.020200
## [430]	train-merror:0.020100
## [431]	train-merror:0.020000
## [432]	train-merror:0.019900
## [433]	train-merror:0.020000
## [434]	train-merror:0.020000
## [435]	train-merror:0.020000
## [436]	train-merror:0.019800
## [437]	train-merror:0.019300
## [438]	train-merror:0.019400
## [439]	train-merror:0.019400
## [440]	train-merror:0.019400
## [441]	train-merror:0.019300
## [442]	train-merror:0.018900
## [443]	train-merror:0.018800
## [444]	train-merror:0.018800
## [445]	train-merror:0.018800
## [446]	train-merror:0.018700
## [447]	train-merror:0.018500
## [448]	train-merror:0.018400
## [449]	train-merror:0.018300
## [450]	train-merror:0.018000
## [451]	train-merror:0.018100
## [452]	train-merror:0.017800
## [453]	train-merror:0.017400
## [454]	train-merror:0.017400
## [455]	train-merror:0.017400
## [456]	train-merror:0.017200
## [457]	train-merror:0.017000
## [458]	train-merror:0.016700
## [459]	train-merror:0.016600
## [460]	train-merror:0.016800
## [461]	train-merror:0.016400
## [462]	train-merror:0.016500
## [463]	train-merror:0.016400
## [464]	train-merror:0.016400
## [465]	train-merror:0.016300
## [466]	train-merror:0.016100
## [467]	train-merror:0.016100
## [468]	train-merror:0.016000
## [469]	train-merror:0.016200
## [470]	train-merror:0.015900
## [471]	train-merror:0.016000
## [472]	train-merror:0.015600
## [473]	train-merror:0.015700
```

```
## [474]    train-merror:0.015600
## [475]    train-merror:0.015500
## [476]    train-merror:0.015300
## [477]    train-merror:0.015000
## [478]    train-merror:0.014900
## [479]    train-merror:0.014800
## [480]    train-merror:0.014700
## [481]    train-merror:0.014600
## [482]    train-merror:0.014700
## [483]    train-merror:0.014500
## [484]    train-merror:0.014400
## [485]    train-merror:0.014400
## [486]    train-merror:0.014300
## [487]    train-merror:0.014000
## [488]    train-merror:0.013800
## [489]    train-merror:0.013700
## [490]    train-merror:0.013600
## [491]    train-merror:0.013600
## [492]    train-merror:0.013300
## [493]    train-merror:0.013300
## [494]    train-merror:0.013300
## [495]    train-merror:0.013000
## [496]    train-merror:0.013100
## [497]    train-merror:0.012600
## [498]    train-merror:0.012300
## [499]    train-merror:0.012300
## [500]    train-merror:0.012200
## [501]    train-merror:0.012100
## [502]    train-merror:0.012100
## [503]    train-merror:0.012100
## [504]    train-merror:0.012200
## [505]    train-merror:0.011900
## [506]    train-merror:0.012000
## [507]    train-merror:0.011900
## [508]    train-merror:0.011800
## [509]    train-merror:0.011900
## [510]    train-merror:0.011600
## [511]    train-merror:0.011500
## [512]    train-merror:0.011500
## [513]    train-merror:0.011100
## [514]    train-merror:0.011100
## [515]    train-merror:0.011000
## [516]    train-merror:0.010900
## [517]    train-merror:0.010700
## [518]    train-merror:0.010900
## [519]    train-merror:0.010700
## [520]    train-merror:0.010500
## [521]    train-merror:0.010500
## [522]    train-merror:0.010400
## [523]    train-merror:0.010300
```

```
## [524]	train-merror:0.010300
## [525]	train-merror:0.010300
## [526]	train-merror:0.010300
## [527]	train-merror:0.010200
## [528]	train-merror:0.010300
## [529]	train-merror:0.010300
## [530]	train-merror:0.010200
## [531]	train-merror:0.010000
## [532]	train-merror:0.010100
## [533]	train-merror:0.009900
## [534]	train-merror:0.009600
## [535]	train-merror:0.009300
## [536]	train-merror:0.009400
## [537]	train-merror:0.009500
## [538]	train-merror:0.009400
## [539]	train-merror:0.009100
## [540]	train-merror:0.009200
## [541]	train-merror:0.009200
## [542]	train-merror:0.009000
## [543]	train-merror:0.009000
## [544]	train-merror:0.008800
## [545]	train-merror:0.008800
## [546]	train-merror:0.008800
## [547]	train-merror:0.008700
## [548]	train-merror:0.008700
## [549]	train-merror:0.008700
## [550]	train-merror:0.008600
## [551]	train-merror:0.008500
## [552]	train-merror:0.008500
## [553]	train-merror:0.008500
## [554]	train-merror:0.008600
## [555]	train-merror:0.008200
## [556]	train-merror:0.008200
## [557]	train-merror:0.008200
## [558]	train-merror:0.007900
## [559]	train-merror:0.007800
## [560]	train-merror:0.007700
## [561]	train-merror:0.007500
## [562]	train-merror:0.007300
## [563]	train-merror:0.007600
## [564]	train-merror:0.007200
## [565]	train-merror:0.007100
## [566]	train-merror:0.007100
## [567]	train-merror:0.007100
## [568]	train-merror:0.007300
## [569]	train-merror:0.006900
## [570]	train-merror:0.006900
## [571]	train-merror:0.006600
## [572]	train-merror:0.006500
## [573]	train-merror:0.006500
```

```
## [574]	train-merror:0.006700
## [575]	train-merror:0.006400
## [576]	train-merror:0.006400
## [577]	train-merror:0.006300
## [578]	train-merror:0.006300
## [579]	train-merror:0.006300
## [580]	train-merror:0.006300
## [581]	train-merror:0.006100
## [582]	train-merror:0.006000
## [583]	train-merror:0.006100
## [584]	train-merror:0.006000
## [585]	train-merror:0.006100
## [586]	train-merror:0.006000
## [587]	train-merror:0.005900
## [588]	train-merror:0.005900
## [589]	train-merror:0.005800
## [590]	train-merror:0.005900
## [591]	train-merror:0.005700
## [592]	train-merror:0.005600
## [593]	train-merror:0.005500
## [594]	train-merror:0.005700
## [595]	train-merror:0.005500
## [596]	train-merror:0.005100
## [597]	train-merror:0.005100
## [598]	train-merror:0.005000
## [599]	train-merror:0.005000
## [600]	train-merror:0.004600
## [601]	train-merror:0.004500
## [602]	train-merror:0.004600
## [603]	train-merror:0.004500
## [604]	train-merror:0.004200
## [605]	train-merror:0.004200
## [606]	train-merror:0.004200
## [607]	train-merror:0.004300
## [608]	train-merror:0.004300
## [609]	train-merror:0.004200
## [610]	train-merror:0.004200
## [611]	train-merror:0.004300
## [612]	train-merror:0.004200
## [613]	train-merror:0.004200
## [614]	train-merror:0.004200
## [615]	train-merror:0.004200
## [616]	train-merror:0.004200
## [617]	train-merror:0.004200
## [618]	train-merror:0.004100
## [619]	train-merror:0.004100
## [620]	train-merror:0.003900
## [621]	train-merror:0.004000
## [622]	train-merror:0.004000
## [623]	train-merror:0.003900
```

```
## [624]	train-merror:0.003900 
## [625]	train-merror:0.003800 
## [626]	train-merror:0.003800 
## [627]	train-merror:0.003800 
## [628]	train-merror:0.003800 
## [629]	train-merror:0.003800 
## [630]	train-merror:0.003700 
## [631]	train-merror:0.003600 
## [632]	train-merror:0.003600 
## [633]	train-merror:0.003500 
## [634]	train-merror:0.003500 
## [635]	train-merror:0.003500 
## [636]	train-merror:0.003500 
## [637]	train-merror:0.003500 
## [638]	train-merror:0.003400 
## [639]	train-merror:0.003400 
## [640]	train-merror:0.003400 
## [641]	train-merror:0.003400 
## [642]	train-merror:0.003300 
## [643]	train-merror:0.003200 
## [644]	train-merror:0.003300 
## [645]	train-merror:0.003200 
## [646]	train-merror:0.003200 
## [647]	train-merror:0.003200 
## [648]	train-merror:0.003200 
## [649]	train-merror:0.003000 
## [650]	train-merror:0.003000 
## [651]	train-merror:0.003000 
## [652]	train-merror:0.003000 
## [653]	train-merror:0.003000 
## [654]	train-merror:0.002900 
## [655]	train-merror:0.003000 
## [656]	train-merror:0.002900 
## [657]	train-merror:0.002900 
## [658]	train-merror:0.002800 
## [659]	train-merror:0.002700 
## [660]	train-merror:0.002700 
## [661]	train-merror:0.002700 
## [662]	train-merror:0.002700 
## [663]	train-merror:0.002800 
## [664]	train-merror:0.002800 
## [665]	train-merror:0.002800 
## [666]	train-merror:0.002700 
## [667]	train-merror:0.002700 
## [668]	train-merror:0.002700 
## [669]	train-merror:0.002700 
## [670]	train-merror:0.002700 
## [671]	train-merror:0.002700 
## [672]	train-merror:0.002700 
## [673]	train-merror:0.002700
```

```
## [674]    train-merror:0.002700
## [675]    train-merror:0.002600
## [676]    train-merror:0.002600
## [677]    train-merror:0.002600
## [678]    train-merror:0.002600
## [679]    train-merror:0.002600
## [680]    train-merror:0.002600
## [681]    train-merror:0.002600
## [682]    train-merror:0.002600
## [683]    train-merror:0.002600
## [684]    train-merror:0.002600
## [685]    train-merror:0.002600
## [686]    train-merror:0.002600
## [687]    train-merror:0.002600
## [688]    train-merror:0.002500
## [689]    train-merror:0.002300
## [690]    train-merror:0.002300
## [691]    train-merror:0.002200
## [692]    train-merror:0.002200
## [693]    train-merror:0.002200
## [694]    train-merror:0.002400
## [695]    train-merror:0.002400
## [696]    train-merror:0.002200
## [697]    train-merror:0.002200
## [698]    train-merror:0.002200
## [699]    train-merror:0.002100
## [700]    train-merror:0.002200
## [701]    train-merror:0.002100
## [702]    train-merror:0.002100
## [703]    train-merror:0.002100
## [704]    train-merror:0.002000
## [705]    train-merror:0.002000
## [706]    train-merror:0.002000
## [707]    train-merror:0.001900
## [708]    train-merror:0.001900
## [709]    train-merror:0.001900
## [710]    train-merror:0.001900
## [711]    train-merror:0.001900
## [712]    train-merror:0.001800
## [713]    train-merror:0.001900
## [714]    train-merror:0.001900
## [715]    train-merror:0.002000
## [716]    train-merror:0.001800
## [717]    train-merror:0.001900
## [718]    train-merror:0.001800
## [719]    train-merror:0.001800
## [720]    train-merror:0.001700
## [721]    train-merror:0.001700
## [722]    train-merror:0.001800
## [723]    train-merror:0.001800
```

```
## [724]	train-merror:0.001700
## [725]	train-merror:0.001700
## [726]	train-merror:0.001500
## [727]	train-merror:0.001600
## [728]	train-merror:0.001500
## [729]	train-merror:0.001400
## [730]	train-merror:0.001400
## [731]	train-merror:0.001400
## [732]	train-merror:0.001400
## [733]	train-merror:0.001400
## [734]	train-merror:0.001300
## [735]	train-merror:0.001400
## [736]	train-merror:0.001400
## [737]	train-merror:0.001400
## [738]	train-merror:0.001300
## [739]	train-merror:0.001300
## [740]	train-merror:0.001300
## [741]	train-merror:0.001300
## [742]	train-merror:0.001300
## [743]	train-merror:0.001300
## [744]	train-merror:0.001300
## [745]	train-merror:0.001300
## [746]	train-merror:0.001300
## [747]	train-merror:0.001300
## [748]	train-merror:0.001300
## [749]	train-merror:0.001400
## [750]	train-merror:0.001300
## [751]	train-merror:0.001300
## [752]	train-merror:0.001300
## [753]	train-merror:0.001300
## [754]	train-merror:0.001400
## [755]	train-merror:0.001400
## [756]	train-merror:0.001400
## [757]	train-merror:0.001300
## [758]	train-merror:0.001200
## [759]	train-merror:0.001100
## [760]	train-merror:0.001100
## [761]	train-merror:0.001100
## [762]	train-merror:0.001100
## [763]	train-merror:0.001100
## [764]	train-merror:0.001000
## [765]	train-merror:0.001000
## [766]	train-merror:0.001000
## [767]	train-merror:0.000900
## [768]	train-merror:0.000800
## [769]	train-merror:0.000800
## [770]	train-merror:0.000800
## [771]	train-merror:0.000800
## [772]	train-merror:0.000800
## [773]	train-merror:0.000800
```

```
## [774]	train-merror:0.000800
## [775]	train-merror:0.000800
## [776]	train-merror:0.000800
## [777]	train-merror:0.000800
## [778]	train-merror:0.000800
## [779]	train-merror:0.000800
## [780]	train-merror:0.000800
## [781]	train-merror:0.000800
## [782]	train-merror:0.000800
## [783]	train-merror:0.000800
## [784]	train-merror:0.000800
## [785]	train-merror:0.000800
## [786]	train-merror:0.000800
## [787]	train-merror:0.000800
## [788]	train-merror:0.000800
## [789]	train-merror:0.000800
## [790]	train-merror:0.000800
## [791]	train-merror:0.000800
## [792]	train-merror:0.000800
## [793]	train-merror:0.000800
## [794]	train-merror:0.000800
## [795]	train-merror:0.000800
## [796]	train-merror:0.000800
## [797]	train-merror:0.000800
## [798]	train-merror:0.000800
## [799]	train-merror:0.000800
## [800]	train-merror:0.000800
## [801]	train-merror:0.000800
## [802]	train-merror:0.000800
## [803]	train-merror:0.000800
## [804]	train-merror:0.000800
## [805]	train-merror:0.000800
## [806]	train-merror:0.000800
## [807]	train-merror:0.000700
## [808]	train-merror:0.000700
## [809]	train-merror:0.000700
## [810]	train-merror:0.000500
## [811]	train-merror:0.000500
## [812]	train-merror:0.000500
## [813]	train-merror:0.000600
## [814]	train-merror:0.000500
## [815]	train-merror:0.000500
## [816]	train-merror:0.000500
## [817]	train-merror:0.000400
## [818]	train-merror:0.000500
## [819]	train-merror:0.000500
## [820]	train-merror:0.000400
## [821]	train-merror:0.000500
## [822]	train-merror:0.000500
## [823]	train-merror:0.000500
```

```
## [824]	train-merror:0.000400
## [825]	train-merror:0.000400
## [826]	train-merror:0.000400
## [827]	train-merror:0.000400
## [828]	train-merror:0.000400
## [829]	train-merror:0.000500
## [830]	train-merror:0.000400
## [831]	train-merror:0.000400
## [832]	train-merror:0.000300
## [833]	train-merror:0.000300
## [834]	train-merror:0.000300
## [835]	train-merror:0.000200
## [836]	train-merror:0.000100
## [837]	train-merror:0.000100
## [838]	train-merror:0.000100
## [839]	train-merror:0.000100
## [840]	train-merror:0.000100
## [841]	train-merror:0.000100
## [842]	train-merror:0.000100
## [843]	train-merror:0.000100
## [844]	train-merror:0.000100
## [845]	train-merror:0.000100
## [846]	train-merror:0.000100
## [847]	train-merror:0.000100
## [848]	train-merror:0.000100
## [849]	train-merror:0.000100
## [850]	train-merror:0.000100
## [851]	train-merror:0.000100
## [852]	train-merror:0.000100
## [853]	train-merror:0.000100
## [854]	train-merror:0.000100
## [855]	train-merror:0.000100
## [856]	train-merror:0.000100
## [857]	train-merror:0.000100
## [858]	train-merror:0.000100
## [859]	train-merror:0.000100
## [860]	train-merror:0.000100
## [861]	train-merror:0.000100
## [862]	train-merror:0.000100
## [863]	train-merror:0.000100
## [864]	train-merror:0.000100
## [865]	train-merror:0.000100
## [866]	train-merror:0.000100
## [867]	train-merror:0.000100
## [868]	train-merror:0.000100
## [869]	train-merror:0.000100
## [870]	train-merror:0.000100
## [871]	train-merror:0.000100
## [872]	train-merror:0.000100
## [873]	train-merror:0.000100
```

```
## [874]	train-merror:0.000100
## [875]	train-merror:0.000100
## [876]	train-merror:0.000100
## [877]	train-merror:0.000100
## [878]	train-merror:0.000100
## [879]	train-merror:0.000100
## [880]	train-merror:0.000100
## [881]	train-merror:0.000100
## [882]	train-merror:0.000100
## [883]	train-merror:0.000100
## [884]	train-merror:0.000100
## [885]	train-merror:0.000100
## [886]	train-merror:0.000100
## [887]	train-merror:0.000100
## [888]	train-merror:0.000100
## [889]	train-merror:0.000100
## [890]	train-merror:0.000100
## [891]	train-merror:0.000100
## [892]	train-merror:0.000100
## [893]	train-merror:0.000100
## [894]	train-merror:0.000100
## [895]	train-merror:0.000100
## [896]	train-merror:0.000100
## [897]	train-merror:0.000100
## [898]	train-merror:0.000100
## [899]	train-merror:0.000100
## [900]	train-merror:0.000100
## [901]	train-merror:0.000100
## [902]	train-merror:0.000100
## [903]	train-merror:0.000100
## [904]	train-merror:0.000100
## [905]	train-merror:0.000100
## [906]	train-merror:0.000100
## [907]	train-merror:0.000100
## [908]	train-merror:0.000100
## [909]	train-merror:0.000100
## [910]	train-merror:0.000100
## [911]	train-merror:0.000100
## [912]	train-merror:0.000100
## [913]	train-merror:0.000100
## [914]	train-merror:0.000100
## [915]	train-merror:0.000100
## [916]	train-merror:0.000100
## [917]	train-merror:0.000000
## [918]	train-merror:0.000000
## [919]	train-merror:0.000000
## [920]	train-merror:0.000000
## [921]	train-merror:0.000000
## [922]	train-merror:0.000000
## [923]	train-merror:0.000000
```

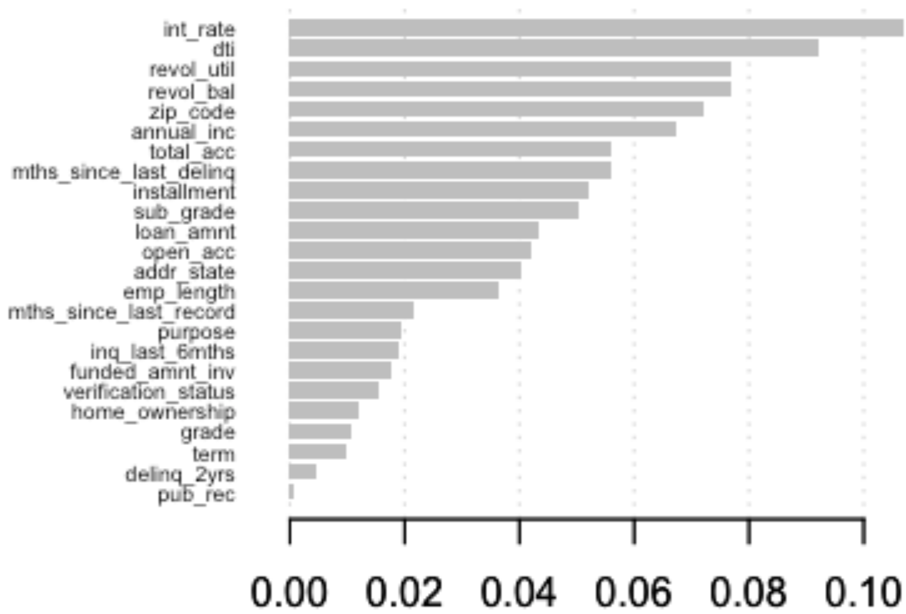
```
## [924]	train-merror:0.000000
## [925]	train-merror:0.000000
## [926]	train-merror:0.000000
## [927]	train-merror:0.000000
## [928]	train-merror:0.000000
## [929]	train-merror:0.000000
## [930]	train-merror:0.000000
## [931]	train-merror:0.000000
## [932]	train-merror:0.000000
## [933]	train-merror:0.000000
## [934]	train-merror:0.000000
## [935]	train-merror:0.000000
## [936]	train-merror:0.000000
## [937]	train-merror:0.000000
## [938]	train-merror:0.000000
## [939]	train-merror:0.000000
## [940]	train-merror:0.000000
## [941]	train-merror:0.000000
## [942]	train-merror:0.000000
## [943]	train-merror:0.000000
## [944]	train-merror:0.000000
## [945]	train-merror:0.000000
## [946]	train-merror:0.000000
## [947]	train-merror:0.000000
## [948]	train-merror:0.000000
## [949]	train-merror:0.000000
## [950]	train-merror:0.000000
## [951]	train-merror:0.000000
## [952]	train-merror:0.000000
## [953]	train-merror:0.000000
## [954]	train-merror:0.000000
## [955]	train-merror:0.000000
## [956]	train-merror:0.000000
## [957]	train-merror:0.000000
## [958]	train-merror:0.000000
## [959]	train-merror:0.000000
## [960]	train-merror:0.000000
## [961]	train-merror:0.000000
## [962]	train-merror:0.000000
## [963]	train-merror:0.000000
## [964]	train-merror:0.000000
## [965]	train-merror:0.000000
## [966]	train-merror:0.000000
## [967]	train-merror:0.000000
## [968]	train-merror:0.000000
## [969]	train-merror:0.000000
## [970]	train-merror:0.000000
## [971]	train-merror:0.000000
## [972]	train-merror:0.000000
## [973]	train-merror:0.000000
```

```
## [974]    train-merror:0.000000
## [975]    train-merror:0.000000
## [976]    train-merror:0.000000
## [977]    train-merror:0.000000
## [978]    train-merror:0.000000
## [979]    train-merror:0.000000
## [980]    train-merror:0.000000
## [981]    train-merror:0.000000
## [982]    train-merror:0.000000
## [983]    train-merror:0.000000
## [984]    train-merror:0.000000
## [985]    train-merror:0.000000
## [986]    train-merror:0.000000
## [987]    train-merror:0.000000
## [988]    train-merror:0.000000
## [989]    train-merror:0.000000
## [990]    train-merror:0.000000
## [991]    train-merror:0.000000
## [992]    train-merror:0.000000
## [993]    train-merror:0.000000
## [994]    train-merror:0.000000
## [995]    train-merror:0.000000
## [996]    train-merror:0.000000
## [997]    train-merror:0.000000
## [998]    train-merror:0.000000
## [999]    train-merror:0.000000
## [1000]   train-merror:0.000000
```

```
#evaluate variable importance
importance <- xgb.importance(feature_names = names(Ytrain[1,-c(1,13)]),
model = xgb)
head(importance,10)
```

```
##                    Feature       Gain      Cover  Frequency
##  1:               int_rate 0.10680964 0.14739700 0.07308296
##  2:                    dti 0.09194389 0.10004324 0.09031660
##  3:             revol_util 0.07688739 0.06468115 0.08342478
##  4:              revol_bal 0.07676184 0.06862776 0.08289950
##  5:               zip_code 0.07202244 0.05835959 0.08209228
##  6:             annual_inc 0.06740165 0.06159518 0.07465061
##  7:              total_acc 0.05595341 0.04897183 0.06167654
##  8: mths_since_last_delinq 0.05592430 0.05053061 0.05992522
##  9:            installment 0.05213834 0.05246922 0.05654541
## 10:              sub_grade 0.05044909 0.06045314 0.02802937
```

```
xgb.plot.importance(importance_matrix = importance)
```



Make prediction on the testing set

```
xgb.pred = predict(xgb,data.matrix(Ytest[,-c(1,13)]))
#calculate AUC
library(pROC)

## Warning: package 'pROC' was built under R version 3.3.2

## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following object is masked from 'package:gmodels':
##
##     ci

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var

multiclass.roc(Ytest$loan_status, xgb.pred, col="black",
         lwd=3, print.auc=TRUE,print.auc.y = 0.0, add=TRUE)
```

```
##
## Call:
## multiclass.roc.default(response = Ytest$loan_status, predictor =
xgb.pred,     col = "black", lwd = 3, print.auc = TRUE, print.auc.y =
0,     add = TRUE)
##
## Data: xgb.pred with 3 levels of Ytest$loan_status: Bad, Good, On
going.
## Multi-class area under the curve: 0.5538
```

The result shows the area under the curve is only 0.5624. It implies that our model is not a strong model in predicting loan status based on the variables selected. There are a few reasons why this is expected.

1. Parameters tuning is not performed yet. We can expect improvements of the model if we choose the optimal parameters, such as learning rate, nrounds, subsamples, maximum depth etc.

2. We used a very small sample relative to the whole dataset(10000 out of 88XXXX). It is reasonable to say that the sample does not capture the rich complexities of the features in the whole dataset, and therefore it has weak predictive power in the testing sample.

3. Not much feature engineering has been done. eg: the variables are not accessed against normality assumption, outliers are not examined.

7. Evaluation

During the study of this dataset, I came across several problems and I think these would benefit future analysis if I have time to explore it further.

1. Correlated variables There are many correlated variables in this dataset and some of them require tedious processing before we can explore the relationships. For example, "url","desc","purpose","title" all contain information of the purpose of the loan, it will be beneficial to extract these information and compare them for anomalies.

2. Text analytics Text analytics can be applied to "desc" and "url" for insights. "desc" contains description by the loaners themselves, and it may reveal similar pattern for loaners who tend to be in a bad status.

3. Anomalous data I detect many anomalous data during graph plotting. For better analysis we can use some R packages to deal with these anomalies.

4. There are so many variables in this dataset. It will be reasonable to remove them by applying relavant knowledge from loan business. It is therefore crucial to understand the process before we can remove any variables and perform feature engineering.

5. The dataset is a big dataset. It is time consuming to perform many analysis and it take up memories exponentially. Maybe we can explore packages like ff, Hadoop and parallel programming to facilitate the process.