download01

data preparation from Kaggle

```
# Lending club is a platform of person-to-person loan. Lending club loan data was downloaded from kaggle web
site, which include a complete record of loan issued between 2007-2015.
# I selected a loan subset from kaggle that are currently under payment. These loan were issued between 2010
-12-01 and 2015-12-01. Their last payment date no later than 2016-01-01.
# This project aims to predict which loan will fail to make next payment.
# A different version of loan data was downloaded from Lending Club website, including updated payment date
records and loan status, which enable me to define labels for next payment prediction.


rm(list=ls())
loank <- read.csv('/Users/fanyang/Documents/lendingclub/lending-club-loan-data/loan.csv', header = TRUE, str
ingsAsFactors = FALSE)
loanT <- loank
table(loank$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
```

```
# select loan subset that is currently under pymnt
loan_request = subset(loank, loan_status == 'Current')
dim(loan_request)
```

```
## [1] 601779     74
```

```
# find out the issued date and last pymnt date
library(zoo)
```

```
##
## Attaching package: 'zoo'
```

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

```
head(loan_request$issue_d)
```

```
## [1] "Dec-2011" "Dec-2011" "Dec-2011" "Dec-2011" "Dec-2011" "Dec-2011"
```

```
as.Date(as.yearmon(loan_request$issue_d[1:5], "%b-%Y"))
```

```
## [1] "2011-12-01" "2011-12-01" "2011-12-01" "2011-12-01" "2011-12-01"
```

```
loan_request$issue_d_1 = as.Date(as.yearmon(loan_request$issue_d, "%b-%Y"))
table(loan_request$issue_d_1)
```

```
##
## 2010-12-01 2011-01-01 2011-02-01 2011-03-01 2011-04-01 2011-05-01
##          7         56        103        142        147        216
## 2011-06-01 2011-07-01 2011-08-01 2011-09-01 2011-10-01 2011-11-01
##        126        152        162        179        200        198
## 2011-12-01 2012-01-01 2012-02-01 2012-03-01 2012-04-01 2012-05-01
##        272        159        143        163        175        213
## 2012-06-01 2012-07-01 2012-08-01 2012-09-01 2012-10-01 2012-11-01
##        223        270        327        374        387        452
## 2012-12-01 2013-01-01 2013-02-01 2013-03-01 2013-04-01 2013-05-01
##        558       1267       2841       3165       3652       4226
## 2013-06-01 2013-07-01 2013-08-01 2013-09-01 2013-10-01 2013-11-01
##       4638       5459       5989       6325       7103       7634
## 2013-12-01 2014-01-01 2014-02-01 2014-03-01 2014-04-01 2014-05-01
##       7965       8656       8996       9821      11714      12047
## 2014-06-01 2014-07-01 2014-08-01 2014-09-01 2014-10-01 2014-11-01
##      11164      19701      12966       7504      28634      19222
## 2014-12-01 2015-01-01 2015-02-01 2015-03-01 2015-04-01 2015-05-01
##       8133      28211      19883      21656      30886      28272
## 2015-06-01 2015-07-01 2015-08-01 2015-09-01 2015-10-01 2015-11-01
##      25747      42542      33890      27278      47208      36710
## 2015-12-01
##      35270
```

```
# Thus, all selected loan were issued between 2010-12-01 and 2015-12-01.
# And their last pymnt date were either 2015-12-01 or 2016-01-01
head(loan_request$last_pymnt_d)
```

```
## [1] "Jan-2016" "Jan-2016" "Jan-2016" "Jan-2016" "Jan-2016" "Jan-2016"
```

```
as.Date(as.yearmon(loan_request$last_pymnt_d[1:5], "%b-%Y"))
```

```
## [1] "2016-01-01" "2016-01-01" "2016-01-01" "2016-01-01" "2016-01-01"
```

```
loan_request$last_pymnt_d_1 = as.Date(as.yearmon(loan_request$last_pymnt_d, "%b-%Y"))
table(loan_request$last_pymnt_d_1)
```

```
##
## 2015-12-01 2016-01-01
##     130624     462829
```

data preparation from Lending Club Statis

```
# Download newly updated loan data from Lending Club website and select loanId that has been included in my
previous search by library(sqldf)

# select loanId issued between 2007 and 2011
d2007_2011T <- read.csv('/Users/fanyang/Documents/lendingclub/original copy/LoanStats3a_securev1.csv', heade
r = TRUE, stringsAsFactors = FALSE, skip = 1)

library("sqldf")
```

```
## Loading required package: gsubfn
```

```
## Loading required package: proto
```

```
## Warning in doTryCatch(return(expr), name, parentenv, handler): unable to load shared object '/Library/Fra
meworks/R.framework/Resources/modules//R_X11.so':
##   dlopen(/Library/Frameworks/R.framework/Resources/modules//R_X11.so, 6): Library not loaded: /opt/X11/li
b/libSM.6.dylib
##   Referenced from: /Library/Frameworks/R.framework/Resources/modules//R_X11.so
##   Reason: image not found
```

```
## Could not load tcltk.  Will use slower R code instead.
```

```
## Loading required package: RSQLite
```

```
d07_11 <- sqldf("SELECT l.*,
d711.id, d711.issue_d, d711.total_pymnt, d711.last_pymnt_d, d711.last_pymnt_amnt,
d711.total_pymnt_inv, d711.total_rec_int, d711.total_rec_late_fee, d711.total_rec_prncp, d711.recoveries, d7
11.collection_recovery_fee, d711.out_prncp, d711.out_prncp_inv, d711.loan_status
FROM loan_request l, d2007_2011T d711
WHERE l.id = d711.id;")

dim(d07_11)
```

```
## [1] 1960   90
```

```
# select loanId issued between 2012 and 2013
d2012_2013 <- read.csv('/Users/fanyang/Documents/lendingclub/original copy/LoanStats3b_securev1.csv', header
= TRUE, stringsAsFactors = FALSE, skip = 1)

d12_13 <- sqldf("SELECT l.*,
d1213.id, d1213.issue_d, d1213.total_pymnt, d1213.last_pymnt_d, d1213.last_pymnt_amnt, d1213.total_pymnt_inv
,d1213.total_rec_int, d1213.total_rec_late_fee, d1213.total_rec_prncp, d1213.recoveries, d1213.collection_re
covery_fee,d1213.out_prncp, d1213.out_prncp_inv, d1213.loan_status
FROM loan_request l, d2012_2013 d1213
WHERE l.id = d1213.id;")

dim(d12_13)
```

```
## [1] 63708    90
```

```
# select loanId issued in 2014
d2014 <- read.csv('/Users/fanyang/Documents/lendingclub/original copy/LoanStats3c_securev1.csv', header = TR
UE, stringsAsFactors = FALSE, skip = 1)


d14 <- sqldf("SELECT l.*,
d2014.id, d2014.issue_d, d2014.total_pymnt, d2014.last_pymnt_d, d2014.last_pymnt_amnt, d2014.total_pymnt_inv
,d2014.total_rec_int, d2014.total_rec_late_fee, d2014.total_rec_prncp, d2014.recoveries, d2014.collection_re
covery_fee,d2014.out_prncp, d2014.out_prncp_inv, d2014.loan_status
FROM loan_request l, d2014
WHERE l.id = d2014.id;")

dim(d14)
```

```
## [1] 158558     90
```

```
# select loanId issued in 2015
d2015 <- read.csv('/Users/fanyang/Documents/lendingclub/original copy/LoanStats3d_securev1.csv', header = TR
UE, stringsAsFactors = FALSE, skip = 1)

d15 <- sqldf("SELECT l.*,
d2015.id, d2015.issue_d, d2015.total_pymnt, d2015.last_pymnt_d, d2015.last_pymnt_amnt, d2015.total_pymnt_inv
,d2015.total_rec_int, d2015.total_rec_late_fee, d2015.total_rec_prncp, d2015.recoveries, d2015.collection_re
covery_fee,d2015.out_prncp, d2015.out_prncp_inv, d2015.loan_status
FROM loan_request l, d2015
WHERE l.id = d2015.id;")

dim(d15)
```

```
## [1] 377553     90
```

```
# select loanId issued in 2016Q1
d2016Q1 <- read.csv('/Users/fanyang/Documents/lendingclub/original copy/LoanStats_securev1_2016Q1.csv', head
er = TRUE, stringsAsFactors = FALSE, skip = 1)

d16Q1 <- sqldf("SELECT l.*,
d2016Q1.id, d2016Q1.issue_d, d2016Q1.total_pymnt, d2016Q1.last_pymnt_d, d2016Q1.last_pymnt_amnt, d2016Q1.tot
al_pymnt_inv,
d2016Q1.total_rec_int, d2016Q1.total_rec_late_fee, d2016Q1.total_rec_prncp, d2016Q1.recoveries, d2016Q1.coll
ection_recovery_fee,
d2016Q1.out_prncp, d2016Q1.out_prncp_inv,  d2016Q1.loan_status
FROM loan_request l, d2016Q1
WHERE l.id = d2016Q1.id;")

dim(d16Q1)
```

```
## [1]  0 90
```

```
# test if my loan_request data include loan issued after 2015
```

```
# combine all collected data into one dataframe
d07_15 <- rbind(d07_11, d12_13, d14, d15)
dim(d07_15)
```

```
## [1] 601779     90
```

```
write.csv(d07_15, file = "/Users/fanyang/Documents/lendingclub/2018_12_21/d2007_2015_col.csv")
```

data pre-processing

```
rm(list=ls())
loan <- read.csv('/Users/fanyang/Documents/lendingclub/2018_12_21/d2007_2015_col.csv', header = TRUE, string
sAsFactors = FALSE)

loans <- loan

dim(loan)
```

```
## [1] 601779     91
```

```
colnames(loans)
```

```
##  [1] "X"                          "id"
##  [3] "member_id"                  "loan_amnt"
##  [5] "funded_amnt"                "funded_amnt_inv"
##  [7] "term"                       "int_rate"
##  [9] "installment"                "grade"
## [11] "sub_grade"                  "emp_title"
## [13] "emp_length"                 "home_ownership"
## [15] "annual_inc"                 "verification_status"
## [17] "issue_d"                    "loan_status"
## [19] "pymnt_plan"                 "url"
## [21] "desc"                       "purpose"
## [23] "title"                      "zip_code"
## [25] "addr_state"                 "dti"
## [27] "delinq_2yrs"                "earliest_cr_line"
## [29] "inq_last_6mths"             "mths_since_last_delinq"
## [31] "mths_since_last_record"     "open_acc"
## [33] "pub_rec"                    "revol_bal"
## [35] "revol_util"                 "total_acc"
## [37] "initial_list_status"        "out_prncp"
## [39] "out_prncp_inv"              "total_pymnt"
## [41] "total_pymnt_inv"            "total_rec_prncp"
## [43] "total_rec_int"              "total_rec_late_fee"
## [45] "recoveries"                 "collection_recovery_fee"
## [47] "last_pymnt_d"               "last_pymnt_amnt"
## [49] "next_pymnt_d"               "last_credit_pull_d"
## [51] "collections_12_mths_ex_med" "mths_since_last_major_derog"
## [53] "policy_code"                "application_type"
## [55] "annual_inc_joint"           "dti_joint"
## [57] "verification_status_joint"  "acc_now_delinq"
## [59] "tot_coll_amt"               "tot_cur_bal"
## [61] "open_acc_6m"                "open_il_6m"
## [63] "open_il_12m"                "open_il_24m"
## [65] "mths_since_rcnt_il"         "total_bal_il"
## [67] "il_util"                    "open_rv_12m"
## [69] "open_rv_24m"                "max_bal_bc"
## [71] "all_util"                   "total_rev_hi_lim"
## [73] "inq_fi"                     "total_cu_tl"
## [75] "inq_last_12m"               "issue_d_1"
## [77] "last_pymnt_d_1"             "id..77"
## [79] "issue_d..78"                "total_pymnt..79"
## [81] "last_pymnt_d..80"           "last_pymnt_amnt..81"
## [83] "total_pymnt_inv..82"        "total_rec_int..83"
## [85] "total_rec_late_fee..84"     "total_rec_prncp..85"
## [87] "recoveries..86"             "collection_recovery_fee..87"
## [89] "out_prncp..88"              "out_prncp_inv..89"
## [91] "loan_status..90"
```

```
# compare previous and subsequent pymnt date
table(loan$last_pymnt_d_1)           # previous pymnt date
```

```
##
## 2015-12-01 2016-01-01
##     130624     462829
```

```
loan$last_pymnt_d..80_1 = as.Date(as.yearmon(loan$last_pymnt_d..80, "%b-%Y"))
table(loan$last_pymnt_d..80_1)       # subsequent pymnt date
```

```
##
## 2015-12-01 2016-01-01 2016-02-01 2016-03-01 2016-04-01 2016-05-01
##        987      10040      15744      18579      15138      15393
## 2016-06-01 2016-07-01 2016-08-01 2016-09-01 2016-10-01 2016-11-01
##      16876      17596      17313      16809      16879      16324
## 2016-12-01 2017-01-01 2017-02-01 2017-03-01 2017-04-01 2017-05-01
##      16439      16555      15791      20353      15616      16364
## 2017-06-01 2017-07-01 2017-08-01 2017-09-01 2017-10-01 2017-11-01
##      16297      16446      16482      12513      15945      14753
## 2017-12-01 2018-01-01 2018-02-01 2018-03-01 2018-04-01 2018-05-01
##      11015      14597      13587      15523      12672      12743
## 2018-06-01 2018-07-01 2018-08-01 2018-09-01 2018-10-01 2018-11-01
##      13277      13425      14224      10296      14172      85010
```

```
# compare loan status
table(loan$loan_status)          # previous loan status
```

```
##
## Current
##  601779
```

```
table(loan$loan_status..90)      # subsequent loan status
```

```
##
##        Charged Off            Current            Default
##              86099              73155                  3
##         Fully Paid    In Grace Period  Late (16-30 days)
##             436972               2108                602
## Late (31-120 days)
##               2840
```

```
# define label: identify loans that probably fail to make next payment
loan$next_pymnt_binary <- with(loan, ifelse(loan_status..90 %in% c('Fully Paid', 'Current'), 0, 1))
sum(loan$next_pymnt_binary)
```

```
## [1] 91652
```

```
# delete variables that are unique for each Id
num.value <- sapply(loan, function(x){return(length(unique(x)))})
names(num.value)[which(num.value == dim(loan)[1])]
```

```
## [1] "X"         "id"        "member_id" "url"       "id..77"
```

```
loan$X = NULL
loan$id = NULL
loan$member_id = NULL
loan$url = NULL
loan$id..77 = NULL
```

```
# delete redundant feature 'dti_joint'
summary(loan$dti)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   12.41   18.24   18.78   24.72 9999.00
```

```
summary(loan$dti_joint)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     3.0    13.2    17.8    18.3    22.6    43.9  601340
```

```
loan$dti = ifelse(!is.na(loan$dti_joint), loan$dti_joint, loan$dti)
loan$dti_joint = NULL
summary(loan$dti)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   12.40   18.24   18.74   24.72   43.86
```

```
# delete redundant feature 'annual_inc_joint'
summary(loan$annual_inc)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0   46000   65000   76153   90000 9500000
```

```
summary(loan$annual_inc_joint)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   17950   75001  100000  107574  131000  410000  601338
```

```
loan$annual_inc = ifelse(!is.na(loan$annual_inc_joint),
                         loan$annual_inc_joint, loan$annual_inc)
loan$annual_inc_joint = NULL
summary(loan$annual_inc)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3000   46000   65000   76189   90000 9500000
```

```
# save data for EDA
write.csv(loan, file = "/Users/fanyang/Documents/lendingclub/2018_12_21/d2007_2015_loan.csv")
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.