Final Project - Predictive Modelling

Amol Gote

08/28/2020

# Dataset Info

1. For this project we will be using the lending loan club dataset, located at <https://www.kaggle.com/wordsforthewise/lending-club>
2. Since this datasize is huge and due to computing constraints, for this project we will be using the data for loans which have originated in 2018. Only approved loans dataset will be in scope for this project.
3. Effective dataset size is of 495,242 observations/loans
4. Attached is the data dictionary file which describes all the variables for this dataset LCDataDictionary.xlsx

# About lending club

The Lending Club is a peer-to-peer lending service (it lends money to customers by matching lenders to borrowers). It enables borrowers to create loan listings on its website by supplying details about themselves and the loans that they would like to request.

# Dataset Attributes

1. Dataset contains various attributes based on which the credit lending decision takes place, this includes Debt to income ratio, annual income, home ownership type (Rented, Mortgage, Owned), employment details.
2. Dataset also contains geographical information like state and zip code. It also has issued loan details like funded Loan Amount, issued month and year, loan Term (36 or 60 months) and interest rate.
3. Post loan has been issued, loan performance needs to be tracked, and this dataset also contains those details loan Status (Current, Paid Off, Charged Off, Delinquent), loan Grades and delinquency details.

# Research Objectives

1. Qualitative - Build a model which will indicate if the loan would ever be delinquent, the delinquency could be of 30, 60, 90 or 120 days.
2. Quantitative - Build a model which will provide the interest rate when the person applies for the loan.
3. Unsupervised - Identify correlation between various credit decisioning variables like fico, dti, annual income,open\_acc,total\_acc, tot\_hi\_cred\_lim, installment, loan\_amnt, etc. using PCA and K-Mean.

Please note that some of the research objective have changed than what were quoted in the project out line document in Week-5,  
a. For quantitative instead of loan amount, now predicting interest rate.  
b. For qualitative predicting only delinquency.

# Feature engineering

Load the whole dataset

lendingClubLoanData <- fread("data/Lending\_club\_dataset.csv", header = TRUE)

Create a origination year field on the dataset, so that all loans which have originated in 2018 can be extracted.

nrow(lendingClubLoanData)

## [1] 2260701

lendingClubLoanData$orig\_year<-substr(lendingClubLoanData$issue\_d,5,8)

lendingClubLoanData\_2018 <- lendingClubLoanData %>%  
 filter(orig\_year == 2018)

lc\_ds\_size <- nrow(lendingClubLoanData\_2018)  
lc\_ds\_size

## [1] 495242

Dataset which will used for this analysis would comprise of 495242 observations.

Create file for all loans which originated in 2018

nrow(lendingClubLoanData\_2018)

## [1] 495242

write.csv(lendingClubLoanData\_2018, file = "data/lending\_club\_loan\_data\_2018\_final.csv", row.names=FALSE)

Load the final data set for analysis

final\_dataset <- fread("data/lending\_club\_loan\_data\_2018\_final.csv", header = TRUE)  
drops <- c("id","member\_id", "url", "desc")  
#final\_dataset[ , !(names(final\_dataset) %in% drops)]

List down all variables in the dataset

names(final\_dataset)

## [1] "id"   
## [2] "member\_id"   
## [3] "loan\_amnt"   
## [4] "funded\_amnt"   
## [5] "funded\_amnt\_inv"   
## [6] "term"   
## [7] "int\_rate"   
## [8] "installment"   
## [9] "grade"   
## [10] "sub\_grade"   
## [11] "emp\_title"   
## [12] "emp\_length"   
## [13] "home\_ownership"   
## [14] "annual\_inc"   
## [15] "verification\_status"   
## [16] "issue\_d"   
## [17] "loan\_status"   
## [18] "pymnt\_plan"   
## [19] "url"   
## [20] "desc"   
## [21] "purpose"   
## [22] "title"   
## [23] "zip\_code"   
## [24] "addr\_state"   
## [25] "dti"   
## [26] "delinq\_2yrs"   
## [27] "earliest\_cr\_line"   
## [28] "fico\_range\_low"   
## [29] "fico\_range\_high"   
## [30] "inq\_last\_6mths"   
## [31] "mths\_since\_last\_delinq"   
## [32] "mths\_since\_last\_record"   
## [33] "open\_acc"   
## [34] "pub\_rec"   
## [35] "revol\_bal"   
## [36] "revol\_util"   
## [37] "total\_acc"   
## [38] "initial\_list\_status"   
## [39] "out\_prncp"   
## [40] "out\_prncp\_inv"   
## [41] "total\_pymnt"   
## [42] "total\_pymnt\_inv"   
## [43] "total\_rec\_prncp"   
## [44] "total\_rec\_int"   
## [45] "total\_rec\_late\_fee"   
## [46] "recoveries"   
## [47] "collection\_recovery\_fee"   
## [48] "last\_pymnt\_d"   
## [49] "last\_pymnt\_amnt"   
## [50] "next\_pymnt\_d"   
## [51] "last\_credit\_pull\_d"   
## [52] "last\_fico\_range\_high"   
## [53] "last\_fico\_range\_low"   
## [54] "collections\_12\_mths\_ex\_med"   
## [55] "mths\_since\_last\_major\_derog"   
## [56] "policy\_code"   
## [57] "application\_type"   
## [58] "annual\_inc\_joint"   
## [59] "dti\_joint"   
## [60] "verification\_status\_joint"   
## [61] "acc\_now\_delinq"   
## [62] "tot\_coll\_amt"   
## [63] "tot\_cur\_bal"   
## [64] "open\_acc\_6m"   
## [65] "open\_act\_il"   
## [66] "open\_il\_12m"   
## [67] "open\_il\_24m"   
## [68] "mths\_since\_rcnt\_il"   
## [69] "total\_bal\_il"   
## [70] "il\_util"   
## [71] "open\_rv\_12m"   
## [72] "open\_rv\_24m"   
## [73] "max\_bal\_bc"   
## [74] "all\_util"   
## [75] "total\_rev\_hi\_lim"   
## [76] "inq\_fi"   
## [77] "total\_cu\_tl"   
## [78] "inq\_last\_12m"   
## [79] "acc\_open\_past\_24mths"   
## [80] "avg\_cur\_bal"   
## [81] "bc\_open\_to\_buy"   
## [82] "bc\_util"   
## [83] "chargeoff\_within\_12\_mths"   
## [84] "delinq\_amnt"   
## [85] "mo\_sin\_old\_il\_acct"   
## [86] "mo\_sin\_old\_rev\_tl\_op"   
## [87] "mo\_sin\_rcnt\_rev\_tl\_op"   
## [88] "mo\_sin\_rcnt\_tl"   
## [89] "mort\_acc"   
## [90] "mths\_since\_recent\_bc"   
## [91] "mths\_since\_recent\_bc\_dlq"   
## [92] "mths\_since\_recent\_inq"   
## [93] "mths\_since\_recent\_revol\_delinq"   
## [94] "num\_accts\_ever\_120\_pd"   
## [95] "num\_actv\_bc\_tl"   
## [96] "num\_actv\_rev\_tl"   
## [97] "num\_bc\_sats"   
## [98] "num\_bc\_tl"   
## [99] "num\_il\_tl"   
## [100] "num\_op\_rev\_tl"   
## [101] "num\_rev\_accts"   
## [102] "num\_rev\_tl\_bal\_gt\_0"   
## [103] "num\_sats"   
## [104] "num\_tl\_120dpd\_2m"   
## [105] "num\_tl\_30dpd"   
## [106] "num\_tl\_90g\_dpd\_24m"   
## [107] "num\_tl\_op\_past\_12m"   
## [108] "pct\_tl\_nvr\_dlq"   
## [109] "percent\_bc\_gt\_75"   
## [110] "pub\_rec\_bankruptcies"   
## [111] "tax\_liens"   
## [112] "tot\_hi\_cred\_lim"   
## [113] "total\_bal\_ex\_mort"   
## [114] "total\_bc\_limit"   
## [115] "total\_il\_high\_credit\_limit"   
## [116] "revol\_bal\_joint"   
## [117] "sec\_app\_fico\_range\_low"   
## [118] "sec\_app\_fico\_range\_high"   
## [119] "sec\_app\_earliest\_cr\_line"   
## [120] "sec\_app\_inq\_last\_6mths"   
## [121] "sec\_app\_mort\_acc"   
## [122] "sec\_app\_open\_acc"   
## [123] "sec\_app\_revol\_util"   
## [124] "sec\_app\_open\_act\_il"   
## [125] "sec\_app\_num\_rev\_accts"   
## [126] "sec\_app\_chargeoff\_within\_12\_mths"   
## [127] "sec\_app\_collections\_12\_mths\_ex\_med"   
## [128] "sec\_app\_mths\_since\_last\_major\_derog"   
## [129] "hardship\_flag"   
## [130] "hardship\_type"   
## [131] "hardship\_reason"   
## [132] "hardship\_status"   
## [133] "deferral\_term"   
## [134] "hardship\_amount"   
## [135] "hardship\_start\_date"   
## [136] "hardship\_end\_date"   
## [137] "payment\_plan\_start\_date"   
## [138] "hardship\_length"   
## [139] "hardship\_dpd"   
## [140] "hardship\_loan\_status"   
## [141] "orig\_projected\_additional\_accrued\_interest"  
## [142] "hardship\_payoff\_balance\_amount"   
## [143] "hardship\_last\_payment\_amount"   
## [144] "disbursement\_method"   
## [145] "debt\_settlement\_flag"   
## [146] "debt\_settlement\_flag\_date"   
## [147] "settlement\_status"   
## [148] "settlement\_date"   
## [149] "settlement\_amount"   
## [150] "settlement\_percentage"   
## [151] "settlement\_term"   
## [152] "orig\_year"

Summary of all the variables in the dataset

summary(final\_dataset)

## id member\_id loan\_amnt funded\_amnt   
## Min. : 70102325 Mode:logical Min. : 1000 Min. : 1000   
## 1st Qu.:131267288 NA's:495242 1st Qu.: 8000 1st Qu.: 8000   
## Median :136195162 Median :14000 Median :14000   
## Mean :136087208 Mean :16025 Mean :16025   
## 3rd Qu.:141042201 3rd Qu.:22000 3rd Qu.:22000   
## Max. :145647287 Max. :40000 Max. :40000   
##   
## funded\_amnt\_inv term int\_rate installment   
## Min. : 725 Length:495242 Min. : 5.31 Min. : 29.76   
## 1st Qu.: 8000 Class :character 1st Qu.: 8.46 1st Qu.: 254.56   
## Median :14000 Mode :character Median :11.80 Median : 386.82   
## Mean :16022 Mean :12.73 Mean : 466.61   
## 3rd Qu.:22000 3rd Qu.:16.01 3rd Qu.: 629.04   
## Max. :40000 Max. :30.99 Max. :1670.15   
##   
## grade sub\_grade emp\_title emp\_length   
## Length:495242 Length:495242 Length:495242 Length:495242   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## home\_ownership annual\_inc verification\_status issue\_d   
## Length:495242 Min. : 0 Length:495242 Length:495242   
## Class :character 1st Qu.: 46000 Class :character Class :character   
## Mode :character Median : 66000 Mode :character Mode :character   
## Mean : 80094   
## 3rd Qu.: 96000   
## Max. :9930475   
##   
## loan\_status pymnt\_plan url desc   
## Length:495242 Length:495242 Length:495242 Mode:logical   
## Class :character Class :character Class :character NA's:495242   
## Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## purpose title zip\_code addr\_state   
## Length:495242 Length:495242 Length:495242 Length:495242   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## dti delinq\_2yrs earliest\_cr\_line fico\_range\_low   
## Min. : 0.00 Min. : 0.0000 Length:495242 Min. :660.0   
## 1st Qu.: 11.43 1st Qu.: 0.0000 Class :character 1st Qu.:680.0   
## Median : 17.71 Median : 0.0000 Mode :character Median :700.0   
## Mean : 19.67 Mean : 0.2293 Mean :706.4   
## 3rd Qu.: 25.03 3rd Qu.: 0.0000 3rd Qu.:725.0   
## Max. :999.00 Max. :58.0000 Max. :845.0   
## NA's :1132   
## fico\_range\_high inq\_last\_6mths mths\_since\_last\_delinq mths\_since\_last\_record  
## Min. :664.0 Min. :0.0000 Min. : 0.00 Min. : 1.0   
## 1st Qu.:684.0 1st Qu.:0.0000 1st Qu.: 19.00 1st Qu.: 69.0   
## Median :704.0 Median :0.0000 Median : 34.00 Median : 87.0   
## Mean :710.4 Mean :0.4422 Mean : 36.89 Mean : 83.3   
## 3rd Qu.:729.0 3rd Qu.:1.0000 3rd Qu.: 53.00 3rd Qu.:102.0   
## Max. :850.0 Max. :5.0000 Max. :226.00 Max. :127.0   
## NA's :276652 NA's :432258   
## open\_acc pub\_rec revol\_bal revol\_util   
## Min. : 0.00 Min. : 0.0000 Min. : 0 Min. : 0.00   
## 1st Qu.: 7.00 1st Qu.: 0.0000 1st Qu.: 5304 1st Qu.: 24.20   
## Median : 10.00 Median : 0.0000 Median : 10832 Median : 42.10   
## Mean : 11.49 Mean : 0.1345 Mean : 16271 Mean : 43.88   
## 3rd Qu.: 14.00 3rd Qu.: 0.0000 3rd Qu.: 19867 3rd Qu.: 62.30   
## Max. :101.00 Max. :52.0000 Max. :2358150 Max. :191.00   
## NA's :592   
## total\_acc initial\_list\_status out\_prncp out\_prncp\_inv   
## Min. : 2.00 Length:495242 Min. : 0 Min. : 0   
## 1st Qu.: 14.00 Class :character 1st Qu.: 4623 1st Qu.: 4622   
## Median : 21.00 Mode :character Median : 9742 Median : 9739   
## Mean : 22.62 Mean :12007 Mean :12005   
## 3rd Qu.: 29.00 3rd Qu.:17801 3rd Qu.:17801   
## Max. :160.00 Max. :40000 Max. :40000   
##   
## total\_pymnt total\_pymnt\_inv total\_rec\_prncp total\_rec\_int   
## Min. : 0 Min. : 0 Min. : 0 Min. : 0.0   
## 1st Qu.: 1758 1st Qu.: 1758 1st Qu.: 1115 1st Qu.: 401.6   
## Median : 3221 Median : 3221 Median : 2075 Median : 814.7   
## Mean : 4925 Mean : 4924 Mean : 3735 Mean : 1179.6   
## 3rd Qu.: 5909 3rd Qu.: 5908 3rd Qu.: 4000 3rd Qu.: 1575.2   
## Max. :51653 Max. :51653 Max. :40000 Max. :12795.0   
##   
## total\_rec\_late\_fee recoveries collection\_recovery\_fee  
## Min. : 0.0000 Min. : 0.00 Min. : 0.000   
## 1st Qu.: 0.0000 1st Qu.: 0.00 1st Qu.: 0.000   
## Median : 0.0000 Median : 0.00 Median : 0.000   
## Mean : 0.5128 Mean : 9.94 Mean : 1.748   
## 3rd Qu.: 0.0000 3rd Qu.: 0.00 3rd Qu.: 0.000   
## Max. :458.5700 Max. :33122.07 Max. :5961.973   
##   
## last\_pymnt\_d last\_pymnt\_amnt next\_pymnt\_d last\_credit\_pull\_d  
## Length:495242 Min. : 0.0 Length:495242 Length:495242   
## Class :character 1st Qu.: 270.7 Class :character Class :character   
## Mode :character Median : 436.9 Mode :character Mode :character   
## Mean : 1599.8   
## 3rd Qu.: 747.9   
## Max. :41353.7   
##   
## last\_fico\_range\_high last\_fico\_range\_low collections\_12\_mths\_ex\_med  
## Min. : 0.0 Min. : 0.0 Min. :0.00000   
## 1st Qu.:679.0 1st Qu.:675.0 1st Qu.:0.00000   
## Median :709.0 Median :705.0 Median :0.00000   
## Mean :709.4 Mean :703.7 Mean :0.01768   
## 3rd Qu.:744.0 3rd Qu.:740.0 3rd Qu.:0.00000   
## Max. :850.0 Max. :845.0 Max. :9.00000   
##   
## mths\_since\_last\_major\_derog policy\_code application\_type annual\_inc\_joint   
## Min. : 0.0 Min. :1 Length:495242 Min. : 5694   
## 1st Qu.: 29.0 1st Qu.:1 Class :character 1st Qu.: 85000   
## Median : 46.0 Median :1 Mode :character Median : 114000   
## Mean : 46.3 Mean :1 Mean : 128342   
## 3rd Qu.: 64.0 3rd Qu.:1 3rd Qu.: 152000   
## Max. :226.0 Max. :1 Max. :7874821   
## NA's :380409 NA's :426257   
## dti\_joint verification\_status\_joint acc\_now\_delinq   
## Min. : 0.0 Length:495242 Min. :0.00e+00   
## 1st Qu.:13.4 Class :character 1st Qu.:0.00e+00   
## Median :18.9 Mode :character Median :0.00e+00   
## Mean :19.4 Mean :5.25e-05   
## 3rd Qu.:25.0 3rd Qu.:0.00e+00   
## Max. :40.0 Max. :1.00e+00   
## NA's :426257   
## tot\_coll\_amt tot\_cur\_bal open\_acc\_6m open\_act\_il   
## Min. : 0 Min. : 0 Min. : 0.0000 Min. : 0.000   
## 1st Qu.: 0 1st Qu.: 26885 1st Qu.: 0.0000 1st Qu.: 1.000   
## Median : 0 Median : 74550 Median : 1.0000 Median : 2.000   
## Mean : 214 Mean : 143960 Mean : 0.8981 Mean : 2.701   
## 3rd Qu.: 0 3rd Qu.: 218044 3rd Qu.: 1.0000 3rd Qu.: 3.000   
## Max. :6214661 Max. :9971659 Max. :15.0000 Max. :56.000   
##   
## open\_il\_12m open\_il\_24m mths\_since\_rcnt\_il total\_bal\_il   
## Min. :0.0000 Min. : 0.000 Min. : 0.0 Min. : 0   
## 1st Qu.:0.0000 1st Qu.: 0.000 1st Qu.: 7.0 1st Qu.: 7851   
## Median :0.0000 Median : 1.000 Median : 13.0 Median : 22452   
## Mean :0.6729 Mean : 1.527 Mean : 21.1 Mean : 35324   
## 3rd Qu.:1.0000 3rd Qu.: 2.000 3rd Qu.: 25.0 3rd Qu.: 45845   
## Max. :8.0000 Max. :20.000 Max. :507.0 Max. :1837038   
## NA's :18410   
## il\_util open\_rv\_12m open\_rv\_24m max\_bal\_bc   
## Min. : 0 Min. : 0.000 Min. : 0.000 Min. : 0   
## 1st Qu.: 53 1st Qu.: 0.000 1st Qu.: 1.000 1st Qu.: 2183   
## Median : 71 Median : 1.000 Median : 2.000 Median : 4441   
## Mean : 68 Mean : 1.228 Mean : 2.624 Mean : 5863   
## 3rd Qu.: 85 3rd Qu.: 2.000 3rd Qu.: 4.000 3rd Qu.: 7772   
## Max. :1000 Max. :26.000 Max. :50.000 Max. :1170668   
## NA's :80824   
## all\_util total\_rev\_hi\_lim inq\_fi total\_cu\_tl   
## Min. : 0.00 Min. : 0 Min. : 0.000 Min. : 0.000   
## 1st Qu.: 40.00 1st Qu.: 16200 1st Qu.: 0.000 1st Qu.: 0.000   
## Median : 55.00 Median : 28700 Median : 1.000 Median : 0.000   
## Mean : 54.09 Mean : 38345 Mean : 1.087 Mean : 1.485   
## 3rd Qu.: 69.00 3rd Qu.: 48700 3rd Qu.: 2.000 3rd Qu.: 2.000   
## Max. :239.00 Max. :2087500 Max. :38.000 Max. :68.000   
## NA's :129   
## inq\_last\_12m acc\_open\_past\_24mths avg\_cur\_bal bc\_open\_to\_buy   
## Min. : 0.000 Min. : 0.000 Min. : 0 Min. : 0   
## 1st Qu.: 0.000 1st Qu.: 2.000 1st Qu.: 2904 1st Qu.: 2867   
## Median : 1.000 Median : 4.000 Median : 7033 Median : 8301   
## Mean : 1.937 Mean : 4.427 Mean : 13709 Mean : 15057   
## 3rd Qu.: 3.000 3rd Qu.: 6.000 3rd Qu.: 19031 3rd Qu.: 19694   
## Max. :67.000 Max. :54.000 Max. :623229 Max. :605996   
## NA's :40 NA's :6588   
## bc\_util chargeoff\_within\_12\_mths delinq\_amnt   
## Min. : 0.00 Min. :0.000000 Min. : 0.00   
## 1st Qu.: 25.95 1st Qu.:0.000000 1st Qu.: 0.00   
## Median : 48.80 Median :0.000000 Median : 0.00   
## Mean : 49.87 Mean :0.006823 Mean : 1.79   
## 3rd Qu.: 74.40 3rd Qu.:0.000000 3rd Qu.: 0.00   
## Max. :201.60 Max. :9.000000 Max. :65000.00   
## NA's :6803   
## mo\_sin\_old\_il\_acct mo\_sin\_old\_rev\_tl\_op mo\_sin\_rcnt\_rev\_tl\_op  
## Min. : 0 Min. : 1.0 Min. : 0.00   
## 1st Qu.: 85 1st Qu.:102.0 1st Qu.: 4.00   
## Median :129 Median :156.0 Median : 9.00   
## Mean :123 Mean :174.6 Mean : 15.13   
## 3rd Qu.:155 3rd Qu.:226.0 3rd Qu.: 19.00   
## Max. :848 Max. :826.0 Max. :502.00   
## NA's :18410   
## mo\_sin\_rcnt\_tl mort\_acc mths\_since\_recent\_bc  
## Min. : 0.000 Min. : 0.000 Min. : 0.00   
## 1st Qu.: 3.000 1st Qu.: 0.000 1st Qu.: 6.00   
## Median : 6.000 Median : 1.000 Median : 15.00   
## Mean : 8.719 Mean : 1.336 Mean : 25.55   
## 3rd Qu.: 11.000 3rd Qu.: 2.000 3rd Qu.: 30.00   
## Max. :382.000 Max. :87.000 Max. :661.00   
## NA's :6198   
## mths\_since\_recent\_bc\_dlq mths\_since\_recent\_inq mths\_since\_recent\_revol\_delinq  
## Min. : 0.0 Min. : 0.00 Min. : 0.0   
## 1st Qu.: 23.0 1st Qu.: 2.00 1st Qu.: 20.0   
## Median : 38.0 Median : 6.00 Median : 35.0   
## Mean : 40.5 Mean : 7.49 Mean : 37.8   
## 3rd Qu.: 57.0 3rd Qu.:11.00 3rd Qu.: 53.0   
## Max. :194.0 Max. :25.00 Max. :190.0   
## NA's :397132 NA's :61305 NA's :352552   
## num\_accts\_ever\_120\_pd num\_actv\_bc\_tl num\_actv\_rev\_tl num\_bc\_sats   
## Min. : 0.0000 Min. : 0.000 Min. : 0.000 Min. : 0.00   
## 1st Qu.: 0.0000 1st Qu.: 2.000 1st Qu.: 3.000 1st Qu.: 3.00   
## Median : 0.0000 Median : 3.000 Median : 5.000 Median : 4.00   
## Mean : 0.4679 Mean : 3.614 Mean : 5.362 Mean : 4.84   
## 3rd Qu.: 0.0000 3rd Qu.: 5.000 3rd Qu.: 7.000 3rd Qu.: 6.00   
## Max. :58.0000 Max. :50.000 Max. :72.000 Max. :69.00   
##   
## num\_bc\_tl num\_il\_tl num\_op\_rev\_tl num\_rev\_accts   
## Min. : 0.000 Min. : 0.000 Min. : 0.000 Min. : 2.00   
## 1st Qu.: 4.000 1st Qu.: 3.000 1st Qu.: 5.000 1st Qu.: 7.00   
## Median : 6.000 Median : 6.000 Median : 7.000 Median : 11.00   
## Mean : 7.093 Mean : 8.169 Mean : 8.163 Mean : 12.92   
## 3rd Qu.: 9.000 3rd Qu.: 11.000 3rd Qu.:10.000 3rd Qu.: 17.00   
## Max. :86.000 Max. :130.000 Max. :91.000 Max. :151.00   
##   
## num\_rev\_tl\_bal\_gt\_0 num\_sats num\_tl\_120dpd\_2m num\_tl\_30dpd   
## Min. : 0.000 Min. : 0.00 Min. :0 Min. :0.00e+00   
## 1st Qu.: 3.000 1st Qu.: 7.00 1st Qu.:0 1st Qu.:0.00e+00   
## Median : 5.000 Median : 10.00 Median :0 Median :0.00e+00   
## Mean : 5.325 Mean : 11.47 Mean :0 Mean :4.64e-05   
## 3rd Qu.: 7.000 3rd Qu.: 14.00 3rd Qu.:0 3rd Qu.:0.00e+00   
## Max. :65.000 Max. :101.00 Max. :0 Max. :1.00e+00   
## NA's :12404   
## num\_tl\_90g\_dpd\_24m num\_tl\_op\_past\_12m pct\_tl\_nvr\_dlq percent\_bc\_gt\_75  
## Min. : 0.00000 Min. : 0.00 Min. : 0.00 Min. : 0.00   
## 1st Qu.: 0.00000 1st Qu.: 1.00 1st Qu.: 92.30 1st Qu.: 0.00   
## Median : 0.00000 Median : 2.00 Median :100.00 Median : 25.00   
## Mean : 0.05993 Mean : 2.03 Mean : 94.58 Mean : 32.89   
## 3rd Qu.: 0.00000 3rd Qu.: 3.00 3rd Qu.:100.00 3rd Qu.: 57.10   
## Max. :58.00000 Max. :26.00 Max. :100.00 Max. :100.00   
## NA's :2 NA's :6596   
## pub\_rec\_bankruptcies tax\_liens tot\_hi\_cred\_lim total\_bal\_ex\_mort  
## Min. :0.0000 Min. : 0.00000 Min. : 0 Min. : 0   
## 1st Qu.:0.0000 1st Qu.: 0.00000 1st Qu.: 52059 1st Qu.: 19530   
## Median :0.0000 Median : 0.00000 Median : 116292 Median : 37556   
## Mean :0.1235 Mean : 0.01092 Mean : 184942 Mean : 51922   
## 3rd Qu.:0.0000 3rd Qu.: 0.00000 3rd Qu.: 269292 3rd Qu.: 66119   
## Max. :7.0000 Max. :52.00000 Max. :9999999 Max. :2622906   
##   
## total\_bc\_limit total\_il\_high\_credit\_limit revol\_bal\_joint   
## Min. : 0 Min. : 0 Min. : 0   
## 1st Qu.: 9800 1st Qu.: 14926 1st Qu.: 15477   
## Median : 19200 Median : 33754 Median : 27350   
## Mean : 26680 Mean : 45688 Mean : 34757   
## 3rd Qu.: 35200 3rd Qu.: 61914 3rd Qu.: 45215   
## Max. :1569000 Max. :2118996 Max. :1110019   
## NA's :426257   
## sec\_app\_fico\_range\_low sec\_app\_fico\_range\_high sec\_app\_earliest\_cr\_line  
## Min. :540.0 Min. :544.0 Length:495242   
## 1st Qu.:645.0 1st Qu.:649.0 Class :character   
## Median :670.0 Median :674.0 Mode :character   
## Mean :671.8 Mean :675.8   
## 3rd Qu.:700.0 3rd Qu.:704.0   
## Max. :845.0 Max. :850.0   
## NA's :426257 NA's :426257   
## sec\_app\_inq\_last\_6mths sec\_app\_mort\_acc sec\_app\_open\_acc sec\_app\_revol\_util  
## Min. :0.0 Min. : 0.0 Min. : 0.0 Min. : 0.0   
## 1st Qu.:0.0 1st Qu.: 0.0 1st Qu.: 7.0 1st Qu.: 38.3   
## Median :0.0 Median : 1.0 Median :10.0 Median : 59.1   
## Mean :0.6 Mean : 1.5 Mean :11.5 Mean : 57.1   
## 3rd Qu.:1.0 3rd Qu.: 2.0 3rd Qu.:15.0 3rd Qu.: 77.8   
## Max. :6.0 Max. :27.0 Max. :67.0 Max. :434.3   
## NA's :426257 NA's :426257 NA's :426257 NA's :427454   
## sec\_app\_open\_act\_il sec\_app\_num\_rev\_accts sec\_app\_chargeoff\_within\_12\_mths  
## Min. : 0 Min. : 0.0 Min. : 0   
## 1st Qu.: 1 1st Qu.: 7.0 1st Qu.: 0   
## Median : 2 Median : 11.0 Median : 0   
## Mean : 3 Mean : 12.5 Mean : 0   
## 3rd Qu.: 4 3rd Qu.: 17.0 3rd Qu.: 0   
## Max. :43 Max. :106.0 Max. :21   
## NA's :426257 NA's :426257 NA's :426257   
## sec\_app\_collections\_12\_mths\_ex\_med sec\_app\_mths\_since\_last\_major\_derog  
## Min. : 0.0 Min. : 0.0   
## 1st Qu.: 0.0 1st Qu.: 17.0   
## Median : 0.0 Median : 36.0   
## Mean : 0.1 Mean : 37.3   
## 3rd Qu.: 0.0 3rd Qu.: 57.0   
## Max. :23.0 Max. :185.0   
## NA's :426257 NA's :472865   
## hardship\_flag hardship\_type hardship\_reason hardship\_status   
## Length:495242 Length:495242 Length:495242 Length:495242   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## deferral\_term hardship\_amount hardship\_start\_date hardship\_end\_date   
## Min. :3 Min. : 6.0 Length:495242 Length:495242   
## 1st Qu.:3 1st Qu.: 87.4 Class :character Class :character   
## Median :3 Median :160.1 Mode :character Mode :character   
## Mean :3 Mean :196.1   
## 3rd Qu.:3 3rd Qu.:271.9   
## Max. :3 Max. :845.2   
## NA's :494874 NA's :494874   
## payment\_plan\_start\_date hardship\_length hardship\_dpd hardship\_loan\_status  
## Length:495242 Min. :3 Min. : 0.0 Length:495242   
## Class :character 1st Qu.:3 1st Qu.: 5.0 Class :character   
## Mode :character Median :3 Median :13.0 Mode :character   
## Mean :3 Mean :12.8   
## 3rd Qu.:3 3rd Qu.:20.0   
## Max. :3 Max. :29.0   
## NA's :494874 NA's :494874   
## orig\_projected\_additional\_accrued\_interest hardship\_payoff\_balance\_amount  
## Min. : 31.5 Min. : 424.1   
## 1st Qu.: 258.7 1st Qu.: 7966.5   
## Median : 479.8 Median :13323.6   
## Mean : 595.4 Mean :15633.3   
## 3rd Qu.: 819.9 3rd Qu.:22590.2   
## Max. :2535.7 Max. :40149.3   
## NA's :494921 NA's :494874   
## hardship\_last\_payment\_amount disbursement\_method debt\_settlement\_flag  
## Min. : 0.1 Length:495242 Length:495242   
## 1st Qu.: 50.9 Class :character Class :character   
## Median : 153.2 Mode :character Mode :character   
## Mean : 215.7   
## 3rd Qu.: 327.4   
## Max. :1159.6   
## NA's :494874   
## debt\_settlement\_flag\_date settlement\_status settlement\_date   
## Length:495242 Length:495242 Length:495242   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## settlement\_amount settlement\_percentage settlement\_term orig\_year   
## Min. : 413.9 Min. :29.9 Min. : 1 Min. :2018   
## 1st Qu.: 3367.1 1st Qu.:45.0 1st Qu.:17 1st Qu.:2018   
## Median : 5626.1 Median :55.0 Median :18 Median :2018   
## Mean : 7145.7 Mean :54.6 Mean :18 Mean :2018   
## 3rd Qu.: 9761.5 3rd Qu.:65.0 3rd Qu.:24 3rd Qu.:2018   
## Max. :28503.0 Max. :80.0 Max. :24 Max. :2018   
## NA's :494762 NA's :494762 NA's :494762

## Variable Selection

1. Dataset contains lot of variables, some of these varibales are associated with secondary applicant, so excluding those variables. Some other variables contain lot of NA values so dropped those variables.

ds\_lc <- final\_dataset[,c("loan\_amnt",  
 "funded\_amnt",  
 "funded\_amnt\_inv",  
 "term",  
 "int\_rate",  
 "installment",  
 "grade",  
 "sub\_grade",  
 "home\_ownership",  
 "annual\_inc",  
 "loan\_status",  
 "dti",  
 "delinq\_2yrs",  
 "fico\_range\_low",  
 "fico\_range\_high",  
 "inq\_last\_6mths",  
 "mths\_since\_last\_delinq",  
 "mths\_since\_last\_record",  
 "open\_acc",  
 "pub\_rec",  
 "revol\_bal",  
 "revol\_util",  
 "total\_acc",  
 "out\_prncp",  
 "out\_prncp\_inv",  
 "total\_pymnt",  
 "total\_pymnt\_inv",  
 "total\_rec\_prncp",  
 "total\_rec\_int",  
 "total\_rec\_late\_fee",  
 "recoveries",  
 "collection\_recovery\_fee",  
 "last\_fico\_range\_high",  
 "last\_fico\_range\_low",  
 "mths\_since\_last\_major\_derog",  
 "acc\_now\_delinq",  
 "avg\_cur\_bal",  
 "inq\_last\_12m", "num\_tl\_30dpd", "num\_tl\_90g\_dpd\_24m", "tot\_hi\_cred\_lim", "num\_rev\_accts"  
 )]

## Clean up the datset

1. Convert all categorical variables to factors
2. Remove the NA values
3. Convert variables to numeric or integer
4. Create a delinquent account flag, loan account becomes deliquent (is\_acct\_delinquent = 1) when loan status turns to either of the following a. Late (16-30 days) b. Late (31-120 days) c. Charged Off d. Default e. In Grace Period

Loan status with following is one which is not delinquent (is\_acct\_delinquent = 0)   
a. Current   
b. Fully Paid

nrow(ds\_lc)

## [1] 495242

ds\_lc <- na.omit(ds\_lc, cols = c("dti", "int\_rate"))  
nrow(ds\_lc)

## [1] 494110

levels(factor(ds\_lc$loan\_status))

## [1] "Charged Off" "Current" "Default"   
## [4] "Fully Paid" "In Grace Period" "Late (16-30 days)"   
## [7] "Late (31-120 days)"

levels(factor(ds\_lc$grade))

## [1] "A" "B" "C" "D" "E" "F" "G"

levels(factor(ds\_lc$sub\_grade))

## [1] "A1" "A2" "A3" "A4" "A5" "B1" "B2" "B3" "B4" "B5" "C1" "C2" "C3" "C4" "C5"  
## [16] "D1" "D2" "D3" "D4" "D5" "E1" "E2" "E3" "E4" "E5" "F1" "F2" "F3" "F4" "F5"  
## [31] "G1" "G2" "G3" "G4" "G5"

levels(factor(ds\_lc$home\_ownership))

## [1] "ANY" "MORTGAGE" "OWN" "RENT"

ds\_lc$term <- as.integer(as.factor(ds\_lc$term))   
ds\_lc$grade <- as.integer(as.factor(ds\_lc$grade))  
ds\_lc$sub\_grade <- as.integer(as.factor(ds\_lc$sub\_grade))  
ds\_lc$home\_ownership <- as.integer(as.factor(ds\_lc$home\_ownership))  
ds\_lc$int\_rate <- as.numeric(ds\_lc$int\_rate)  
ds\_lc$loan\_status <- as.integer(as.factor(ds\_lc$loan\_status))

cor(ds\_lc)

## loan\_amnt funded\_amnt funded\_amnt\_inv  
## loan\_amnt 1.000000000 1.000000000 0.999996431  
## funded\_amnt 1.000000000 1.000000000 0.999996431  
## funded\_amnt\_inv 0.999996431 0.999996431 1.000000000  
## term 0.382920212 0.382920212 0.383125468  
## int\_rate 0.003852733 0.003852733 0.003953011  
## installment 0.943022509 0.943022509 0.942953979  
## grade 0.007639237 0.007639237 0.007730099  
## sub\_grade 0.003261095 0.003261095 0.003355532  
## home\_ownership -0.156677785 -0.156677785 -0.156718890  
## annual\_inc 0.235938716 0.235938716 0.235942416  
## loan\_status -0.016390465 -0.016390465 -0.016388719  
## dti 0.046349677 0.046349677 0.046348146  
## delinq\_2yrs -0.017199481 -0.017199481 -0.017220019  
## fico\_range\_low 0.098274031 0.098274031 0.098349805  
## fico\_range\_high 0.098272857 0.098272857 0.098348632  
## inq\_last\_6mths -0.016122851 -0.016122851 -0.016150955  
## mths\_since\_last\_delinq NA NA NA  
## mths\_since\_last\_record NA NA NA  
## open\_acc 0.167820213 0.167820213 0.167804824  
## pub\_rec -0.043718920 -0.043718920 -0.043755256  
## revol\_bal 0.303135377 0.303135377 0.303138333  
## revol\_util NA NA NA  
## total\_acc 0.182344359 0.182344359 0.182336495  
## out\_prncp 0.833688171 0.833688171 0.833726070  
## out\_prncp\_inv 0.833696709 0.833696709 0.833739941  
## total\_pymnt 0.503599930 0.503599930 0.503532033  
## total\_pymnt\_inv 0.503657020 0.503657020 0.503593731  
## total\_rec\_prncp 0.394606358 0.394606358 0.394529897  
## total\_rec\_int 0.651241262 0.651241262 0.651258290  
## total\_rec\_late\_fee 0.049554213 0.049554213 0.049548150  
## recoveries 0.039263939 0.039263939 0.039277142  
## collection\_recovery\_fee 0.039122821 0.039122821 0.039135789  
## last\_fico\_range\_high 0.139337824 0.139337824 0.139370350  
## last\_fico\_range\_low 0.112018908 0.112018908 0.112043035  
## mths\_since\_last\_major\_derog NA NA NA  
## acc\_now\_delinq 0.002585038 0.002585038 0.002587481  
## avg\_cur\_bal NA NA NA  
## inq\_last\_12m 0.014312391 0.014312391 0.014294214  
## num\_tl\_30dpd 0.002281560 0.002281560 0.002283855  
## num\_tl\_90g\_dpd\_24m -0.019319259 -0.019319259 -0.019334430  
## tot\_hi\_cred\_lim 0.306986363 0.306986363 0.307018054  
## num\_rev\_accts 0.151574978 0.151574978 0.151552012  
## term int\_rate installment grade  
## loan\_amnt 0.382920212 0.003852733 0.943022509 0.007639237  
## funded\_amnt 0.382920212 0.003852733 0.943022509 0.007639237  
## funded\_amnt\_inv 0.383125468 0.003953011 0.942953979 0.007730099  
## term 1.000000000 0.315873119 0.126833134 0.317162573  
## int\_rate 0.315873119 1.000000000 0.045627846 0.970650942  
## installment 0.126833134 0.045627846 1.000000000 0.043225133  
## grade 0.317162573 0.970650942 0.043225133 1.000000000  
## sub\_grade 0.319772499 0.991921262 0.041078723 0.971497931  
## home\_ownership -0.105698102 0.076672572 -0.121842351 0.073167063  
## annual\_inc 0.038244945 -0.063022576 0.230372246 -0.061654515  
## loan\_status -0.012869648 0.066049488 -0.003936516 0.070778493  
## dti 0.042599609 0.108866984 0.052845661 0.104493955  
## delinq\_2yrs -0.018733536 0.104924407 0.002580986 0.105277093  
## fico\_range\_low 0.026621452 -0.444131200 0.031200628 -0.445972966  
## fico\_range\_high 0.026620931 -0.444125375 0.031200221 -0.445966261  
## inq\_last\_6mths 0.004523212 0.118095260 -0.002392947 0.116427626  
## mths\_since\_last\_delinq NA NA NA NA  
## mths\_since\_last\_record NA NA NA NA  
## open\_acc 0.065090967 -0.021399766 0.158881395 -0.021914670  
## pub\_rec 0.001024193 0.042710237 -0.037138644 0.048914936  
## revol\_bal 0.074986410 -0.021959008 0.294927862 -0.021542199  
## revol\_util NA NA NA NA  
## total\_acc 0.088376674 -0.044985807 0.160721339 -0.043563239  
## out\_prncp 0.400473074 0.022804997 0.755170398 0.015454447  
## out\_prncp\_inv 0.400625130 0.022877332 0.755132050 0.015522361  
## total\_pymnt 0.086325134 0.016828958 0.525196558 0.037658206  
## total\_pymnt\_inv 0.086475063 0.016900280 0.525211385 0.037722080  
## total\_rec\_prncp -0.004656810 -0.072107643 0.424086743 -0.052744615  
## total\_rec\_int 0.442989252 0.410399237 0.621183739 0.423245590  
## total\_rec\_late\_fee 0.010277979 0.049057407 0.059775068 0.050555783  
## recoveries 0.032448532 0.043509232 0.038725644 0.044924543  
## collection\_recovery\_fee 0.032263467 0.042321929 0.038341875 0.043776267  
## last\_fico\_range\_high 0.021666730 -0.405871937 0.087427902 -0.409761571  
## last\_fico\_range\_low 0.015103595 -0.343083135 0.068997809 -0.346999768  
## mths\_since\_last\_major\_derog NA NA NA NA  
## acc\_now\_delinq 0.004308905 0.003262334 0.001811403 0.003391450  
## avg\_cur\_bal NA NA NA NA  
## inq\_last\_12m 0.034148361 0.123783615 0.021788226 0.122993832  
## num\_tl\_30dpd 0.003879019 0.003342816 0.001693180 0.003411623  
## num\_tl\_90g\_dpd\_24m -0.009061182 0.053309127 -0.010692559 0.054599713  
## tot\_hi\_cred\_lim 0.106249415 -0.118911443 0.272687554 -0.116377823  
## num\_rev\_accts 0.044782512 -0.070099920 0.140900805 -0.067119423  
## sub\_grade home\_ownership annual\_inc  
## loan\_amnt 0.003261095 -0.156677785 0.2359387158  
## funded\_amnt 0.003261095 -0.156677785 0.2359387158  
## funded\_amnt\_inv 0.003355532 -0.156718890 0.2359424163  
## term 0.319772499 -0.105698102 0.0382449454  
## int\_rate 0.991921262 0.076672572 -0.0630225764  
## installment 0.041078723 -0.121842351 0.2303722461  
## grade 0.971497931 0.073167063 -0.0616545152  
## sub\_grade 1.000000000 0.080295606 -0.0671319853  
## home\_ownership 0.080295606 1.000000000 -0.1184635377  
## annual\_inc -0.067131985 -0.118463538 1.0000000000  
## loan\_status 0.073515991 -0.013121088 0.0009358584  
## dti 0.109993559 -0.051518077 -0.1052857082  
## delinq\_2yrs 0.108261874 -0.050270263 0.0413779496  
## fico\_range\_low -0.460446298 -0.095890360 0.0180188910  
## fico\_range\_high -0.460440587 -0.095892682 0.0180185365  
## inq\_last\_6mths 0.122450351 -0.022157249 0.0346032241  
## mths\_since\_last\_delinq NA NA NA  
## mths\_since\_last\_record NA NA NA  
## open\_acc -0.023157065 -0.153695856 0.1435176618  
## pub\_rec 0.047855284 -0.010093267 -0.0122393305  
## revol\_bal -0.025138176 -0.161546851 0.2655227511  
## revol\_util NA NA NA  
## total\_acc -0.047576674 -0.238039827 0.1652009564  
## out\_prncp 0.011455899 -0.127512832 0.1901004364  
## out\_prncp\_inv 0.011526088 -0.127547895 0.1901050232  
## total\_pymnt 0.037166893 -0.089287524 0.1256627543  
## total\_pymnt\_inv 0.037233452 -0.089321251 0.1256825785  
## total\_rec\_prncp -0.055048499 -0.080790819 0.1116953873  
## total\_rec\_int 0.431305163 -0.067500837 0.1023674781  
## total\_rec\_late\_fee 0.052196916 -0.003540954 0.0161302822  
## recoveries 0.046097406 0.006369627 0.0020853917  
## collection\_recovery\_fee 0.044960357 0.006387752 0.0020665567  
## last\_fico\_range\_high -0.424990864 -0.089059154 0.0116669230  
## last\_fico\_range\_low -0.359534889 -0.085630068 0.0122406117  
## mths\_since\_last\_major\_derog NA NA NA  
## acc\_now\_delinq 0.004319267 -0.002884071 0.0022103967  
## avg\_cur\_bal NA NA NA  
## inq\_last\_12m 0.129298598 -0.118031996 0.0652556146  
## num\_tl\_30dpd 0.004345513 -0.002834661 0.0021593505  
## num\_tl\_90g\_dpd\_24m 0.055119728 -0.004363646 0.0078984880  
## tot\_hi\_cred\_lim -0.126109654 -0.515406238 0.3525239483  
## num\_rev\_accts -0.072150375 -0.137934044 0.1031635158  
## loan\_status dti delinq\_2yrs  
## loan\_amnt -0.0163904647 0.0463496769 -0.0171994812  
## funded\_amnt -0.0163904647 0.0463496769 -0.0171994812  
## funded\_amnt\_inv -0.0163887190 0.0463481460 -0.0172200190  
## term -0.0128696480 0.0425996090 -0.0187335359  
## int\_rate 0.0660494876 0.1088669839 0.1049244068  
## installment -0.0039365160 0.0528456611 0.0025809856  
## grade 0.0707784926 0.1044939546 0.1052770934  
## sub\_grade 0.0735159911 0.1099935589 0.1082618744  
## home\_ownership -0.0131210883 -0.0515180769 -0.0502702633  
## annual\_inc 0.0009358584 -0.1052857082 0.0413779496  
## loan\_status 1.0000000000 -0.0109872516 0.0102141488  
## dti -0.0109872516 1.0000000000 -0.0150975436  
## delinq\_2yrs 0.0102141488 -0.0150975436 1.0000000000  
## fico\_range\_low -0.0040314464 -0.0069413954 -0.1701543335  
## fico\_range\_high -0.0040321216 -0.0069444200 -0.1701510849  
## inq\_last\_6mths 0.0472363760 -0.0087823748 0.0123729038  
## mths\_since\_last\_delinq NA NA NA  
## mths\_since\_last\_record NA NA NA  
## open\_acc -0.0041501266 0.1487021268 0.0288303802  
## pub\_rec 0.0161570330 -0.0071320523 -0.0411543703  
## revol\_bal -0.0242396970 0.1013595912 -0.0196930899  
## revol\_util NA NA NA  
## total\_acc 0.0284192530 0.1298543781 0.0998368294  
## out\_prncp -0.2071870633 0.0549809079 -0.0157018713  
## out\_prncp\_inv -0.2071438782 0.0549788886 -0.0157177821  
## total\_pymnt 0.3828427605 0.0093071615 -0.0005063478  
## total\_pymnt\_inv 0.3828231008 0.0093086295 -0.0005194844  
## total\_rec\_prncp 0.4185412745 -0.0083290007 -0.0069666975  
## total\_rec\_int -0.0382425937 0.0852225697 0.0292715207  
## total\_rec\_late\_fee 0.1247694431 -0.0038166645 0.0222086630  
## recoveries -0.0731974135 -0.0027747091 0.0027054680  
## collection\_recovery\_fee -0.0724438596 -0.0026845570 0.0028620101  
## last\_fico\_range\_high -0.0782368560 0.0036098758 -0.1345538219  
## last\_fico\_range\_low -0.0548607169 0.0078471314 -0.1040541545  
## mths\_since\_last\_major\_derog NA NA NA  
## acc\_now\_delinq 0.0027540386 0.0001528656 0.0187593389  
## avg\_cur\_bal NA NA NA  
## inq\_last\_12m 0.0594053230 0.0256614528 0.0265063066  
## num\_tl\_30dpd 0.0032050388 0.0001340608 0.0190237430  
## num\_tl\_90g\_dpd\_24m 0.0053381647 -0.0118091268 0.6613762046  
## tot\_hi\_cred\_lim 0.0119832679 0.0553516667 0.0551921414  
## num\_rev\_accts 0.0179007051 0.0822274175 0.0713271331  
## fico\_range\_low fico\_range\_high inq\_last\_6mths  
## loan\_amnt 0.098274031 0.098272857 -0.016122851  
## funded\_amnt 0.098274031 0.098272857 -0.016122851  
## funded\_amnt\_inv 0.098349805 0.098348632 -0.016150955  
## term 0.026621452 0.026620931 0.004523212  
## int\_rate -0.444131200 -0.444125375 0.118095260  
## installment 0.031200628 0.031200221 -0.002392947  
## grade -0.445972966 -0.445966261 0.116427626  
## sub\_grade -0.460446298 -0.460440587 0.122450351  
## home\_ownership -0.095890360 -0.095892682 -0.022157249  
## annual\_inc 0.018018891 0.018018537 0.034603224  
## loan\_status -0.004031446 -0.004032122 0.047236376  
## dti -0.006941395 -0.006944420 -0.008782375  
## delinq\_2yrs -0.170154333 -0.170151085 0.012372904  
## fico\_range\_low 1.000000000 0.999999871 -0.062192863  
## fico\_range\_high 0.999999871 1.000000000 -0.062194999  
## inq\_last\_6mths -0.062192863 -0.062194999 1.000000000  
## mths\_since\_last\_delinq NA NA NA  
## mths\_since\_last\_record NA NA NA  
## open\_acc 0.012021954 0.012020035 0.164187338  
## pub\_rec -0.184051967 -0.184048602 0.062155756  
## revol\_bal -0.041008215 -0.041008810 -0.016190588  
## revol\_util NA NA NA  
## total\_acc 0.034670400 0.034669192 0.156985413  
## out\_prncp 0.060329202 0.060328764 -0.035275313  
## out\_prncp\_inv 0.060386666 0.060386228 -0.035291464  
## total\_pymnt 0.056836538 0.056835391 0.018632101  
## total\_pymnt\_inv 0.056895890 0.056894744 0.018608717  
## total\_rec\_prncp 0.086714822 0.086713235 0.013425625  
## total\_rec\_int -0.120702734 -0.120700990 0.025836182  
## total\_rec\_late\_fee -0.016651086 -0.016649736 0.016956911  
## recoveries -0.004793327 -0.004793630 0.022502093  
## collection\_recovery\_fee -0.004420242 -0.004420554 0.022132159  
## last\_fico\_range\_high 0.554239033 0.554238275 -0.122643044  
## last\_fico\_range\_low 0.448696854 0.448696278 -0.108650483  
## mths\_since\_last\_major\_derog NA NA NA  
## acc\_now\_delinq -0.006627401 -0.006627230 0.006733042  
## avg\_cur\_bal NA NA NA  
## inq\_last\_12m -0.101708460 -0.101710646 0.462566380  
## num\_tl\_30dpd -0.006806362 -0.006806181 0.005246655  
## num\_tl\_90g\_dpd\_24m -0.101673474 -0.101671145 0.015905696  
## tot\_hi\_cred\_lim 0.154338775 0.154339115 0.042407648  
## num\_rev\_accts 0.028998546 0.028999704 0.161304719  
## mths\_since\_last\_delinq mths\_since\_last\_record  
## loan\_amnt NA NA  
## funded\_amnt NA NA  
## funded\_amnt\_inv NA NA  
## term NA NA  
## int\_rate NA NA  
## installment NA NA  
## grade NA NA  
## sub\_grade NA NA  
## home\_ownership NA NA  
## annual\_inc NA NA  
## loan\_status NA NA  
## dti NA NA  
## delinq\_2yrs NA NA  
## fico\_range\_low NA NA  
## fico\_range\_high NA NA  
## inq\_last\_6mths NA NA  
## mths\_since\_last\_delinq 1 NA  
## mths\_since\_last\_record NA 1  
## open\_acc NA NA  
## pub\_rec NA NA  
## revol\_bal NA NA  
## revol\_util NA NA  
## total\_acc NA NA  
## out\_prncp NA NA  
## out\_prncp\_inv NA NA  
## total\_pymnt NA NA  
## total\_pymnt\_inv NA NA  
## total\_rec\_prncp NA NA  
## total\_rec\_int NA NA  
## total\_rec\_late\_fee NA NA  
## recoveries NA NA  
## collection\_recovery\_fee NA NA  
## last\_fico\_range\_high NA NA  
## last\_fico\_range\_low NA NA  
## mths\_since\_last\_major\_derog NA NA  
## acc\_now\_delinq NA NA  
## avg\_cur\_bal NA NA  
## inq\_last\_12m NA NA  
## num\_tl\_30dpd NA NA  
## num\_tl\_90g\_dpd\_24m NA NA  
## tot\_hi\_cred\_lim NA NA  
## num\_rev\_accts NA NA  
## open\_acc pub\_rec revol\_bal revol\_util  
## loan\_amnt 0.167820213 -4.371892e-02 0.3031353773 NA  
## funded\_amnt 0.167820213 -4.371892e-02 0.3031353773 NA  
## funded\_amnt\_inv 0.167804824 -4.375526e-02 0.3031383327 NA  
## term 0.065090967 1.024193e-03 0.0749864102 NA  
## int\_rate -0.021399766 4.271024e-02 -0.0219590079 NA  
## installment 0.158881395 -3.713864e-02 0.2949278617 NA  
## grade -0.021914670 4.891494e-02 -0.0215421988 NA  
## sub\_grade -0.023157065 4.785528e-02 -0.0251381760 NA  
## home\_ownership -0.153695856 -1.009327e-02 -0.1615468511 NA  
## annual\_inc 0.143517662 -1.223933e-02 0.2655227511 NA  
## loan\_status -0.004150127 1.615703e-02 -0.0242396970 NA  
## dti 0.148702127 -7.132052e-03 0.1013595912 NA  
## delinq\_2yrs 0.028830380 -4.115437e-02 -0.0196930899 NA  
## fico\_range\_low 0.012021954 -1.840520e-01 -0.0410082150 NA  
## fico\_range\_high 0.012020035 -1.840486e-01 -0.0410088101 NA  
## inq\_last\_6mths 0.164187338 6.215576e-02 -0.0161905876 NA  
## mths\_since\_last\_delinq NA NA NA NA  
## mths\_since\_last\_record NA NA NA NA  
## open\_acc 1.000000000 9.374260e-03 0.2591593358 NA  
## pub\_rec 0.009374260 1.000000e+00 -0.0758159549 NA  
## revol\_bal 0.259159336 -7.581595e-02 1.0000000000 NA  
## revol\_util NA NA NA 1  
## total\_acc 0.746398102 2.547150e-03 0.2309505887 NA  
## out\_prncp 0.144114220 -4.697160e-02 0.2655340819 NA  
## out\_prncp\_inv 0.144104420 -4.699343e-02 0.2655370548 NA  
## total\_pymnt 0.083412824 -1.155224e-03 0.1369187052 NA  
## total\_pymnt\_inv 0.083413578 -1.187024e-03 0.1369401124 NA  
## total\_rec\_prncp 0.067731492 -4.015431e-03 0.1091035866 NA  
## total\_rec\_int 0.098613012 1.213911e-02 0.1706493186 NA  
## total\_rec\_late\_fee -0.010710187 -3.085977e-03 0.0055731399 NA  
## recoveries -0.003134903 4.709848e-03 -0.0008746645 NA  
## collection\_recovery\_fee -0.003200564 4.761926e-03 -0.0008297269 NA  
## last\_fico\_range\_high 0.024910200 -1.045679e-01 0.0265839257 NA  
## last\_fico\_range\_low 0.030089056 -8.231886e-02 0.0273805795 NA  
## mths\_since\_last\_major\_derog NA NA NA NA  
## acc\_now\_delinq 0.001458211 -3.616169e-04 0.0043697065 NA  
## avg\_cur\_bal NA NA NA NA  
## inq\_last\_12m 0.189774845 7.843465e-02 -0.0268563474 NA  
## num\_tl\_30dpd 0.001774289 -7.338182e-05 0.0029268879 NA  
## num\_tl\_90g\_dpd\_24m -0.001887868 -2.101857e-02 -0.0276327652 NA  
## tot\_hi\_cred\_lim 0.327264787 -6.110352e-02 0.4584005412 NA  
## num\_rev\_accts 0.720558787 2.948937e-02 0.2341479214 NA  
## total\_acc out\_prncp out\_prncp\_inv  
## loan\_amnt 0.182344359 0.8336881713 0.8336967093  
## funded\_amnt 0.182344359 0.8336881713 0.8336967093  
## funded\_amnt\_inv 0.182336495 0.8337260697 0.8337399408  
## term 0.088376674 0.4004730743 0.4006251297  
## int\_rate -0.044985807 0.0228049967 0.0228773323  
## installment 0.160721339 0.7551703979 0.7551320503  
## grade -0.043563239 0.0154544469 0.0155223608  
## sub\_grade -0.047576674 0.0114558993 0.0115260876  
## home\_ownership -0.238039827 -0.1275128322 -0.1275478946  
## annual\_inc 0.165200956 0.1901004364 0.1901050232  
## loan\_status 0.028419253 -0.2071870633 -0.2071438782  
## dti 0.129854378 0.0549809079 0.0549788886  
## delinq\_2yrs 0.099836829 -0.0157018713 -0.0157177821  
## fico\_range\_low 0.034670400 0.0603292016 0.0603866656  
## fico\_range\_high 0.034669192 0.0603287636 0.0603862275  
## inq\_last\_6mths 0.156985413 -0.0352753132 -0.0352914643  
## mths\_since\_last\_delinq NA NA NA  
## mths\_since\_last\_record NA NA NA  
## open\_acc 0.746398102 0.1441142201 0.1441044198  
## pub\_rec 0.002547150 -0.0469716001 -0.0469934286  
## revol\_bal 0.230950589 0.2655340819 0.2655370548  
## revol\_util NA NA NA  
## total\_acc 1.000000000 0.1368120534 0.1368129619  
## out\_prncp 0.136812053 1.0000000000 0.9999977116  
## out\_prncp\_inv 0.136812962 0.9999977116 1.0000000000  
## total\_pymnt 0.118083740 0.0046486205 0.0046603115  
## total\_pymnt\_inv 0.118087348 0.0047159266 0.0047301931  
## total\_rec\_prncp 0.106228851 -0.1186719933 -0.1186633906  
## total\_rec\_int 0.091122039 0.5852788190 0.5852947341  
## total\_rec\_late\_fee -0.006598512 0.0235820773 0.0235804478  
## recoveries -0.002183928 -0.0649772280 -0.0649658336  
## collection\_recovery\_fee -0.002120730 -0.0643082993 -0.0642970222  
## last\_fico\_range\_high 0.047273251 0.1600303205 0.1600492265  
## last\_fico\_range\_low 0.053348858 0.1537705394 0.1537809058  
## mths\_since\_last\_major\_derog NA NA NA  
## acc\_now\_delinq 0.001601258 0.0011905812 0.0011923941  
## avg\_cur\_bal NA NA NA  
## inq\_last\_12m 0.220514209 -0.0151437164 -0.0151504147  
## num\_tl\_30dpd 0.001994190 0.0007663134 0.0007680118  
## num\_tl\_90g\_dpd\_24m 0.055517446 -0.0166346840 -0.0166475560  
## tot\_hi\_cred\_lim 0.388567775 0.2439154976 0.2439442955  
## num\_rev\_accts 0.771322840 0.1131072098 0.1130959721  
## total\_pymnt total\_pymnt\_inv total\_rec\_prncp  
## loan\_amnt 0.5035999302 0.5036570201 3.946064e-01  
## funded\_amnt 0.5035999302 0.5036570201 3.946064e-01  
## funded\_amnt\_inv 0.5035320325 0.5035937311 3.945299e-01  
## term 0.0863251344 0.0864750628 -4.656810e-03  
## int\_rate 0.0168289582 0.0169002800 -7.210764e-02  
## installment 0.5251965581 0.5252113848 4.240867e-01  
## grade 0.0376582059 0.0377220799 -5.274461e-02  
## sub\_grade 0.0371668930 0.0372334524 -5.504850e-02  
## home\_ownership -0.0892875235 -0.0893212507 -8.079082e-02  
## annual\_inc 0.1256627543 0.1256825785 1.116954e-01  
## loan\_status 0.3828427605 0.3828231008 4.185413e-01  
## dti 0.0093071615 0.0093086295 -8.329001e-03  
## delinq\_2yrs -0.0005063478 -0.0005194844 -6.966698e-03  
## fico\_range\_low 0.0568365379 0.0568958904 8.671482e-02  
## fico\_range\_high 0.0568353914 0.0568947438 8.671324e-02  
## inq\_last\_6mths 0.0186321007 0.0186087168 1.342563e-02  
## mths\_since\_last\_delinq NA NA NA  
## mths\_since\_last\_record NA NA NA  
## open\_acc 0.0834128243 0.0834135778 6.773149e-02  
## pub\_rec -0.0011552240 -0.0011870244 -4.015431e-03  
## revol\_bal 0.1369187052 0.1369401124 1.091036e-01  
## revol\_util NA NA NA  
## total\_acc 0.1180837401 0.1180873481 1.062289e-01  
## out\_prncp 0.0046486205 0.0047159266 -1.186720e-01  
## out\_prncp\_inv 0.0046603115 0.0047301931 -1.186634e-01  
## total\_pymnt 1.0000000000 0.9999977379 9.796335e-01  
## total\_pymnt\_inv 0.9999977379 1.0000000000 9.796182e-01  
## total\_rec\_prncp 0.9796334868 0.9796182278 1.000000e+00  
## total\_rec\_int 0.3949109534 0.3949688282 2.055864e-01  
## total\_rec\_late\_fee 0.0406593060 0.0406586664 2.313606e-02  
## recoveries 0.0079373189 0.0079478183 -2.639359e-02  
## collection\_recovery\_fee 0.0080785014 0.0080888879 -2.608477e-02  
## last\_fico\_range\_high 0.0982162636 0.0982446138 1.260119e-01  
## last\_fico\_range\_low 0.0893757256 0.0893962591 1.128106e-01  
## mths\_since\_last\_major\_derog NA NA NA  
## acc\_now\_delinq 0.0025415251 0.0025435166 -3.323112e-05  
## avg\_cur\_bal NA NA NA  
## inq\_last\_12m 0.0466694780 0.0466548130 3.738382e-02  
## num\_tl\_30dpd 0.0023314647 0.0023333275 -8.733205e-05  
## num\_tl\_90g\_dpd\_24m -0.0066079789 -0.0066167025 -8.598754e-03  
## tot\_hi\_cred\_lim 0.1739818571 0.1740209404 1.583238e-01  
## num\_rev\_accts 0.0956047864 0.0955968217 8.796602e-02  
## total\_rec\_int total\_rec\_late\_fee recoveries  
## loan\_amnt 0.651241262 0.049554213 0.0392639388  
## funded\_amnt 0.651241262 0.049554213 0.0392639388  
## funded\_amnt\_inv 0.651258290 0.049548150 0.0392771416  
## term 0.442989252 0.010277979 0.0324485321  
## int\_rate 0.410399237 0.049057407 0.0435092320  
## installment 0.621183739 0.059775068 0.0387256437  
## grade 0.423245590 0.050555783 0.0449245427  
## sub\_grade 0.431305163 0.052196916 0.0460974063  
## home\_ownership -0.067500837 -0.003540954 0.0063696271  
## annual\_inc 0.102367478 0.016130282 0.0020853917  
## loan\_status -0.038242594 0.124769443 -0.0731974135  
## dti 0.085222570 -0.003816665 -0.0027747091  
## delinq\_2yrs 0.029271521 0.022208663 0.0027054680  
## fico\_range\_low -0.120702734 -0.016651086 -0.0047933270  
## fico\_range\_high -0.120700990 -0.016649736 -0.0047936304  
## inq\_last\_6mths 0.025836182 0.016956911 0.0225020928  
## mths\_since\_last\_delinq NA NA NA  
## mths\_since\_last\_record NA NA NA  
## open\_acc 0.098613012 -0.010710187 -0.0031349027  
## pub\_rec 0.012139110 -0.003085977 0.0047098483  
## revol\_bal 0.170649319 0.005573140 -0.0008746645  
## revol\_util NA NA NA  
## total\_acc 0.091122039 -0.006598512 -0.0021839278  
## out\_prncp 0.585278819 0.023582077 -0.0649772280  
## out\_prncp\_inv 0.585294734 0.023580448 -0.0649658336  
## total\_pymnt 0.394910953 0.040659306 0.0079373189  
## total\_pymnt\_inv 0.394968828 0.040658666 0.0079478183  
## total\_rec\_prncp 0.205586355 0.023136062 -0.0263935912  
## total\_rec\_int 1.000000000 0.085183551 -0.0108265011  
## total\_rec\_late\_fee 0.085183551 1.000000000 0.0191620433  
## recoveries -0.010826501 0.019162043 1.0000000000  
## collection\_recovery\_fee -0.010701043 0.019247306 0.9950056570  
## last\_fico\_range\_high -0.074905753 -0.096040528 -0.1393440923  
## last\_fico\_range\_low -0.052866463 -0.084289698 -0.1655039722  
## mths\_since\_last\_major\_derog NA NA NA  
## acc\_now\_delinq 0.010252575 0.005826567 0.0141583980  
## avg\_cur\_bal NA NA NA  
## inq\_last\_12m 0.054191849 0.020657188 0.0169163537  
## num\_tl\_30dpd 0.009304432 0.005447726 0.0150988682  
## num\_tl\_90g\_dpd\_24m 0.006796814 0.005458839 0.0024714283  
## tot\_hi\_cred\_lim 0.125699770 0.016736497 -0.0031025983  
## num\_rev\_accts 0.064444155 -0.011869097 -0.0002349269  
## collection\_recovery\_fee last\_fico\_range\_high  
## loan\_amnt 0.0391228212 0.139337824  
## funded\_amnt 0.0391228212 0.139337824  
## funded\_amnt\_inv 0.0391357888 0.139370350  
## term 0.0322634669 0.021666730  
## int\_rate 0.0423219292 -0.405871937  
## installment 0.0383418752 0.087427902  
## grade 0.0437762671 -0.409761571  
## sub\_grade 0.0449603574 -0.424990864  
## home\_ownership 0.0063877515 -0.089059154  
## annual\_inc 0.0020665567 0.011666923  
## loan\_status -0.0724438596 -0.078236856  
## dti -0.0026845570 0.003609876  
## delinq\_2yrs 0.0028620101 -0.134553822  
## fico\_range\_low -0.0044202421 0.554239033  
## fico\_range\_high -0.0044205541 0.554238275  
## inq\_last\_6mths 0.0221321591 -0.122643044  
## mths\_since\_last\_delinq NA NA  
## mths\_since\_last\_record NA NA  
## open\_acc -0.0032005637 0.024910200  
## pub\_rec 0.0047619264 -0.104567944  
## revol\_bal -0.0008297269 0.026583926  
## revol\_util NA NA  
## total\_acc -0.0021207302 0.047273251  
## out\_prncp -0.0643082993 0.160030320  
## out\_prncp\_inv -0.0642970222 0.160049227  
## total\_pymnt 0.0080785014 0.098216264  
## total\_pymnt\_inv 0.0080888879 0.098244614  
## total\_rec\_prncp -0.0260847663 0.126011907  
## total\_rec\_int -0.0107010427 -0.074905753  
## total\_rec\_late\_fee 0.0192473062 -0.096040528  
## recoveries 0.9950056570 -0.139344092  
## collection\_recovery\_fee 1.0000000000 -0.137493527  
## last\_fico\_range\_high -0.1374935265 1.000000000  
## last\_fico\_range\_low -0.1629165589 0.903453483  
## mths\_since\_last\_major\_derog NA NA  
## acc\_now\_delinq 0.0143597196 -0.007039362  
## avg\_cur\_bal NA NA  
## inq\_last\_12m 0.0164488731 -0.140877531  
## num\_tl\_30dpd 0.0153124488 -0.006502756  
## num\_tl\_90g\_dpd\_24m 0.0025185123 -0.076209291  
## tot\_hi\_cred\_lim -0.0030646041 0.113451659  
## num\_rev\_accts -0.0002601800 0.046356075  
## last\_fico\_range\_low mths\_since\_last\_major\_derog  
## loan\_amnt 0.112018908 NA  
## funded\_amnt 0.112018908 NA  
## funded\_amnt\_inv 0.112043035 NA  
## term 0.015103595 NA  
## int\_rate -0.343083135 NA  
## installment 0.068997809 NA  
## grade -0.346999768 NA  
## sub\_grade -0.359534889 NA  
## home\_ownership -0.085630068 NA  
## annual\_inc 0.012240612 NA  
## loan\_status -0.054860717 NA  
## dti 0.007847131 NA  
## delinq\_2yrs -0.104054155 NA  
## fico\_range\_low 0.448696854 NA  
## fico\_range\_high 0.448696278 NA  
## inq\_last\_6mths -0.108650483 NA  
## mths\_since\_last\_delinq NA NA  
## mths\_since\_last\_record NA NA  
## open\_acc 0.030089056 NA  
## pub\_rec -0.082318865 NA  
## revol\_bal 0.027380579 NA  
## revol\_util NA NA  
## total\_acc 0.053348858 NA  
## out\_prncp 0.153770539 NA  
## out\_prncp\_inv 0.153780906 NA  
## total\_pymnt 0.089375726 NA  
## total\_pymnt\_inv 0.089396259 NA  
## total\_rec\_prncp 0.112810597 NA  
## total\_rec\_int -0.052866463 NA  
## total\_rec\_late\_fee -0.084289698 NA  
## recoveries -0.165503972 NA  
## collection\_recovery\_fee -0.162916559 NA  
## last\_fico\_range\_high 0.903453483 NA  
## last\_fico\_range\_low 1.000000000 NA  
## mths\_since\_last\_major\_derog NA 1  
## acc\_now\_delinq -0.005444696 NA  
## avg\_cur\_bal NA NA  
## inq\_last\_12m -0.117767140 NA  
## num\_tl\_30dpd -0.005026549 NA  
## num\_tl\_90g\_dpd\_24m -0.059997743 NA  
## tot\_hi\_cred\_lim 0.104035423 NA  
## num\_rev\_accts 0.046084682 NA  
## acc\_now\_delinq avg\_cur\_bal inq\_last\_12m  
## loan\_amnt 2.585038e-03 NA 0.0143123912  
## funded\_amnt 2.585038e-03 NA 0.0143123912  
## funded\_amnt\_inv 2.587481e-03 NA 0.0142942144  
## term 4.308905e-03 NA 0.0341483606  
## int\_rate 3.262334e-03 NA 0.1237836153  
## installment 1.811403e-03 NA 0.0217882259  
## grade 3.391450e-03 NA 0.1229938317  
## sub\_grade 4.319267e-03 NA 0.1292985976  
## home\_ownership -2.884071e-03 NA -0.1180319955  
## annual\_inc 2.210397e-03 NA 0.0652556146  
## loan\_status 2.754039e-03 NA 0.0594053230  
## dti 1.528656e-04 NA 0.0256614528  
## delinq\_2yrs 1.875934e-02 NA 0.0265063066  
## fico\_range\_low -6.627401e-03 NA -0.1017084599  
## fico\_range\_high -6.627230e-03 NA -0.1017106463  
## inq\_last\_6mths 6.733042e-03 NA 0.4625663802  
## mths\_since\_last\_delinq NA NA NA  
## mths\_since\_last\_record NA NA NA  
## open\_acc 1.458211e-03 NA 0.1897748449  
## pub\_rec -3.616169e-04 NA 0.0784346472  
## revol\_bal 4.369707e-03 NA -0.0268563474  
## revol\_util NA NA NA  
## total\_acc 1.601258e-03 NA 0.2205142085  
## out\_prncp 1.190581e-03 NA -0.0151437164  
## out\_prncp\_inv 1.192394e-03 NA -0.0151504147  
## total\_pymnt 2.541525e-03 NA 0.0466694780  
## total\_pymnt\_inv 2.543517e-03 NA 0.0466548130  
## total\_rec\_prncp -3.323112e-05 NA 0.0373838167  
## total\_rec\_int 1.025258e-02 NA 0.0541918493  
## total\_rec\_late\_fee 5.826567e-03 NA 0.0206571882  
## recoveries 1.415840e-02 NA 0.0169163537  
## collection\_recovery\_fee 1.435972e-02 NA 0.0164488731  
## last\_fico\_range\_high -7.039362e-03 NA -0.1408775314  
## last\_fico\_range\_low -5.444696e-03 NA -0.1177671395  
## mths\_since\_last\_major\_derog NA NA NA  
## acc\_now\_delinq 1.000000e+00 NA 0.0006622403  
## avg\_cur\_bal NA 1 NA  
## inq\_last\_12m 6.622403e-04 NA 1.0000000000  
## num\_tl\_30dpd 9.405371e-01 NA 0.0004295357  
## num\_tl\_90g\_dpd\_24m 6.921225e-03 NA 0.0172775975  
## tot\_hi\_cred\_lim 3.241562e-03 NA 0.1381390343  
## num\_rev\_accts 2.012808e-03 NA 0.1614614819  
## num\_tl\_30dpd num\_tl\_90g\_dpd\_24m tot\_hi\_cred\_lim  
## loan\_amnt 2.281560e-03 -0.019319259 0.306986363  
## funded\_amnt 2.281560e-03 -0.019319259 0.306986363  
## funded\_amnt\_inv 2.283855e-03 -0.019334430 0.307018054  
## term 3.879019e-03 -0.009061182 0.106249415  
## int\_rate 3.342816e-03 0.053309127 -0.118911443  
## installment 1.693180e-03 -0.010692559 0.272687554  
## grade 3.411623e-03 0.054599713 -0.116377823  
## sub\_grade 4.345513e-03 0.055119728 -0.126109654  
## home\_ownership -2.834661e-03 -0.004363646 -0.515406238  
## annual\_inc 2.159350e-03 0.007898488 0.352523948  
## loan\_status 3.205039e-03 0.005338165 0.011983268  
## dti 1.340608e-04 -0.011809127 0.055351667  
## delinq\_2yrs 1.902374e-02 0.661376205 0.055192141  
## fico\_range\_low -6.806362e-03 -0.101673474 0.154338775  
## fico\_range\_high -6.806181e-03 -0.101671145 0.154339115  
## inq\_last\_6mths 5.246655e-03 0.015905696 0.042407648  
## mths\_since\_last\_delinq NA NA NA  
## mths\_since\_last\_record NA NA NA  
## open\_acc 1.774289e-03 -0.001887868 0.327264787  
## pub\_rec -7.338182e-05 -0.021018570 -0.061103523  
## revol\_bal 2.926888e-03 -0.027632765 0.458400541  
## revol\_util NA NA NA  
## total\_acc 1.994190e-03 0.055517446 0.388567775  
## out\_prncp 7.663134e-04 -0.016634684 0.243915498  
## out\_prncp\_inv 7.680118e-04 -0.016647556 0.243944295  
## total\_pymnt 2.331465e-03 -0.006607979 0.173981857  
## total\_pymnt\_inv 2.333328e-03 -0.006616702 0.174020940  
## total\_rec\_prncp -8.733205e-05 -0.008598754 0.158323784  
## total\_rec\_int 9.304432e-03 0.006796814 0.125699770  
## total\_rec\_late\_fee 5.447726e-03 0.005458839 0.016736497  
## recoveries 1.509887e-02 0.002471428 -0.003102598  
## collection\_recovery\_fee 1.531245e-02 0.002518512 -0.003064604  
## last\_fico\_range\_high -6.502756e-03 -0.076209291 0.113451659  
## last\_fico\_range\_low -5.026549e-03 -0.059997743 0.104035423  
## mths\_since\_last\_major\_derog NA NA NA  
## acc\_now\_delinq 9.405371e-01 0.006921225 0.003241562  
## avg\_cur\_bal NA NA NA  
## inq\_last\_12m 4.295357e-04 0.017277598 0.138139034  
## num\_tl\_30dpd 1.000000e+00 0.005370969 0.003537508  
## num\_tl\_90g\_dpd\_24m 5.370969e-03 1.000000000 0.004441019  
## tot\_hi\_cred\_lim 3.537508e-03 0.004441019 1.000000000  
## num\_rev\_accts 2.206563e-03 0.024040048 0.218833601  
## num\_rev\_accts  
## loan\_amnt 0.1515749782  
## funded\_amnt 0.1515749782  
## funded\_amnt\_inv 0.1515520116  
## term 0.0447825122  
## int\_rate -0.0700999201  
## installment 0.1409008046  
## grade -0.0671194226  
## sub\_grade -0.0721503745  
## home\_ownership -0.1379340442  
## annual\_inc 0.1031635158  
## loan\_status 0.0179007051  
## dti 0.0822274175  
## delinq\_2yrs 0.0713271331  
## fico\_range\_low 0.0289985455  
## fico\_range\_high 0.0289997036  
## inq\_last\_6mths 0.1613047186  
## mths\_since\_last\_delinq NA  
## mths\_since\_last\_record NA  
## open\_acc 0.7205587870  
## pub\_rec 0.0294893720  
## revol\_bal 0.2341479214  
## revol\_util NA  
## total\_acc 0.7713228396  
## out\_prncp 0.1131072098  
## out\_prncp\_inv 0.1130959721  
## total\_pymnt 0.0956047864  
## total\_pymnt\_inv 0.0955968217  
## total\_rec\_prncp 0.0879660205  
## total\_rec\_int 0.0644441551  
## total\_rec\_late\_fee -0.0118690967  
## recoveries -0.0002349269  
## collection\_recovery\_fee -0.0002601800  
## last\_fico\_range\_high 0.0463560755  
## last\_fico\_range\_low 0.0460846818  
## mths\_since\_last\_major\_derog NA  
## acc\_now\_delinq 0.0020128077  
## avg\_cur\_bal NA  
## inq\_last\_12m 0.1614614819  
## num\_tl\_30dpd 0.0022065630  
## num\_tl\_90g\_dpd\_24m 0.0240400480  
## tot\_hi\_cred\_lim 0.2188336012  
## num\_rev\_accts 1.0000000000

Based on the above correlation matrix, following variables are highly collinear 1. loan\_amnt, funded\_amnt, funded\_amnt\_inv 2. int\_rate, grade, sub\_grade 3. fico\_range\_low, fico\_range\_high 4. last\_fico\_range\_high, last\_fico\_range\_low 5. total\_pymnt, total\_pymnt\_inv, total\_rec\_prncp

Dropping following variables due to NA values mths\_since\_last\_delinq, mths\_since\_last\_record, mths\_since\_last\_major\_derog, avg\_cur\_bal

ds\_lc = subset(ds\_lc, select = -c(funded\_amnt, funded\_amnt\_inv, grade, sub\_grade, fico\_range\_low, last\_fico\_range\_low, total\_pymnt\_inv, total\_rec\_prncp, mths\_since\_last\_delinq, mths\_since\_last\_record, mths\_since\_last\_major\_derog, avg\_cur\_bal, revol\_util))

ds\_lc$is\_acct\_delinquent[ds\_lc$loan\_status == 1] <- 1  
ds\_lc$is\_acct\_delinquent[ds\_lc$loan\_status == 2] <- 0  
ds\_lc$is\_acct\_delinquent[ds\_lc$loan\_status == 3] <- 1  
ds\_lc$is\_acct\_delinquent[ds\_lc$loan\_status == 4] <- 0  
ds\_lc$is\_acct\_delinquent[ds\_lc$loan\_status == 5] <- 1  
ds\_lc$is\_acct\_delinquent[ds\_lc$loan\_status == 6] <- 0  
ds\_lc$is\_acct\_delinquent[ds\_lc$loan\_status == 7] <- 1  
ds\_lc$is\_acct\_delinquent <- factor(ds\_lc$is\_acct\_delinquent)  
levels(factor(ds\_lc$is\_acct\_delinquent))

## [1] "0" "1"

ds\_lc = subset(ds\_lc, select = -c(loan\_status))  
ds\_lc\_after\_correlation <- ds\_lc

## Regular subset variable selection test

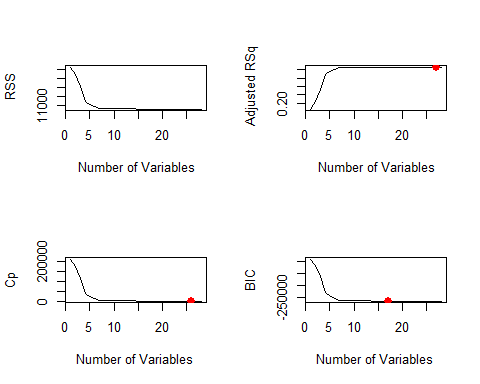
regular\_subset\_full <- regsubsets(is\_acct\_delinquent ~ ., ds\_lc, nvmax = 40)  
regular\_subset\_full\_summary <- summary(regular\_subset\_full)  
regular\_subset\_full\_summary

## Subset selection object  
## Call: regsubsets.formula(is\_acct\_delinquent ~ ., ds\_lc, nvmax = 40)  
## 28 Variables (and intercept)  
## Forced in Forced out  
## loan\_amnt FALSE FALSE  
## term FALSE FALSE  
## int\_rate FALSE FALSE  
## installment FALSE FALSE  
## home\_ownership FALSE FALSE  
## annual\_inc FALSE FALSE  
## dti FALSE FALSE  
## delinq\_2yrs FALSE FALSE  
## fico\_range\_high FALSE FALSE  
## inq\_last\_6mths FALSE FALSE  
## open\_acc FALSE FALSE  
## pub\_rec FALSE FALSE  
## revol\_bal FALSE FALSE  
## total\_acc FALSE FALSE  
## out\_prncp FALSE FALSE  
## out\_prncp\_inv FALSE FALSE  
## total\_pymnt FALSE FALSE  
## total\_rec\_int FALSE FALSE  
## total\_rec\_late\_fee FALSE FALSE  
## recoveries FALSE FALSE  
## collection\_recovery\_fee FALSE FALSE  
## last\_fico\_range\_high FALSE FALSE  
## acc\_now\_delinq FALSE FALSE  
## inq\_last\_12m FALSE FALSE  
## num\_tl\_30dpd FALSE FALSE  
## num\_tl\_90g\_dpd\_24m FALSE FALSE  
## tot\_hi\_cred\_lim FALSE FALSE  
## num\_rev\_accts FALSE FALSE  
## 1 subsets of each size up to 28  
## Selection Algorithm: exhaustive  
## loan\_amnt term int\_rate installment home\_ownership annual\_inc dti  
## 1 ( 1 ) " " " " " " " " " " " " " "  
## 2 ( 1 ) " " " " " " " " " " " " " "  
## 3 ( 1 ) "\*" " " " " " " " " " " " "  
## 4 ( 1 ) "\*" " " " " " " " " " " " "  
## 5 ( 1 ) "\*" " " " " " " " " " " " "  
## 6 ( 1 ) "\*" " " " " " " " " " " " "  
## 7 ( 1 ) "\*" " " " " " " " " " " " "  
## 8 ( 1 ) "\*" " " "\*" " " " " " " " "  
## 9 ( 1 ) "\*" " " "\*" " " " " " " " "  
## 10 ( 1 ) "\*" " " "\*" " " " " "\*" " "  
## 11 ( 1 ) "\*" " " "\*" " " " " "\*" " "  
## 12 ( 1 ) "\*" "\*" "\*" "\*" " " "\*" " "  
## 13 ( 1 ) "\*" " " "\*" " " " " "\*" " "  
## 14 ( 1 ) "\*" "\*" "\*" "\*" " " "\*" " "  
## 15 ( 1 ) "\*" "\*" "\*" "\*" " " "\*" " "  
## 16 ( 1 ) "\*" "\*" "\*" "\*" " " "\*" " "  
## 17 ( 1 ) "\*" "\*" "\*" "\*" " " "\*" "\*"  
## 18 ( 1 ) "\*" "\*" "\*" "\*" " " "\*" "\*"  
## 19 ( 1 ) "\*" "\*" "\*" "\*" " " "\*" "\*"  
## 20 ( 1 ) "\*" "\*" "\*" "\*" " " "\*" "\*"  
## 21 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*"  
## 22 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*"  
## 23 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*"  
## 24 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*"  
## 25 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*"  
## 26 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*"  
## 27 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*"  
## 28 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*"  
## delinq\_2yrs fico\_range\_high inq\_last\_6mths open\_acc pub\_rec revol\_bal  
## 1 ( 1 ) " " " " " " " " " " " "   
## 2 ( 1 ) " " "\*" " " " " " " " "   
## 3 ( 1 ) " " " " " " " " " " " "   
## 4 ( 1 ) " " " " " " " " " " " "   
## 5 ( 1 ) " " " " " " " " " " " "   
## 6 ( 1 ) " " "\*" " " " " " " " "   
## 7 ( 1 ) " " "\*" " " " " " " " "   
## 8 ( 1 ) " " "\*" " " " " " " " "   
## 9 ( 1 ) " " "\*" " " " " " " " "   
## 10 ( 1 ) " " "\*" " " " " " " " "   
## 11 ( 1 ) "\*" "\*" " " " " " " " "   
## 12 ( 1 ) " " "\*" " " " " " " " "   
## 13 ( 1 ) "\*" "\*" " " "\*" " " " "   
## 14 ( 1 ) " " "\*" " " "\*" " " " "   
## 15 ( 1 ) "\*" "\*" " " "\*" " " " "   
## 16 ( 1 ) "\*" "\*" " " "\*" " " " "   
## 17 ( 1 ) "\*" "\*" " " "\*" " " " "   
## 18 ( 1 ) "\*" "\*" " " "\*" " " " "   
## 19 ( 1 ) "\*" "\*" " " "\*" "\*" " "   
## 20 ( 1 ) "\*" "\*" " " "\*" "\*" " "   
## 21 ( 1 ) "\*" "\*" " " "\*" "\*" " "   
## 22 ( 1 ) "\*" "\*" " " "\*" "\*" " "   
## 23 ( 1 ) "\*" "\*" " " "\*" "\*" " "   
## 24 ( 1 ) "\*" "\*" " " "\*" "\*" "\*"   
## 25 ( 1 ) "\*" "\*" " " "\*" "\*" "\*"   
## 26 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*"   
## 27 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*"   
## 28 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*"   
## total\_acc out\_prncp out\_prncp\_inv total\_pymnt total\_rec\_int  
## 1 ( 1 ) " " " " " " " " " "   
## 2 ( 1 ) " " " " " " " " " "   
## 3 ( 1 ) " " "\*" " " "\*" " "   
## 4 ( 1 ) " " "\*" " " "\*" " "   
## 5 ( 1 ) " " "\*" " " "\*" "\*"   
## 6 ( 1 ) " " "\*" " " "\*" "\*"   
## 7 ( 1 ) " " "\*" " " "\*" "\*"   
## 8 ( 1 ) " " "\*" " " "\*" "\*"   
## 9 ( 1 ) " " "\*" " " "\*" "\*"   
## 10 ( 1 ) " " "\*" " " "\*" "\*"   
## 11 ( 1 ) " " "\*" " " "\*" "\*"   
## 12 ( 1 ) " " "\*" " " "\*" "\*"   
## 13 ( 1 ) " " "\*" " " "\*" "\*"   
## 14 ( 1 ) " " "\*" " " "\*" "\*"   
## 15 ( 1 ) " " "\*" " " "\*" "\*"   
## 16 ( 1 ) " " "\*" " " "\*" "\*"   
## 17 ( 1 ) " " "\*" " " "\*" "\*"   
## 18 ( 1 ) " " "\*" " " "\*" "\*"   
## 19 ( 1 ) " " "\*" " " "\*" "\*"   
## 20 ( 1 ) "\*" "\*" " " "\*" "\*"   
## 21 ( 1 ) "\*" "\*" " " "\*" "\*"   
## 22 ( 1 ) "\*" "\*" " " "\*" "\*"   
## 23 ( 1 ) "\*" "\*" "\*" "\*" "\*"   
## 24 ( 1 ) "\*" "\*" "\*" "\*" "\*"   
## 25 ( 1 ) "\*" "\*" "\*" "\*" "\*"   
## 26 ( 1 ) "\*" "\*" "\*" "\*" "\*"   
## 27 ( 1 ) "\*" "\*" "\*" "\*" "\*"   
## 28 ( 1 ) "\*" "\*" "\*" "\*" "\*"   
## total\_rec\_late\_fee recoveries collection\_recovery\_fee  
## 1 ( 1 ) " " " " " "   
## 2 ( 1 ) " " " " " "   
## 3 ( 1 ) " " " " " "   
## 4 ( 1 ) " " " " " "   
## 5 ( 1 ) " " " " " "   
## 6 ( 1 ) " " " " " "   
## 7 ( 1 ) "\*" " " " "   
## 8 ( 1 ) "\*" " " " "   
## 9 ( 1 ) "\*" "\*" " "   
## 10 ( 1 ) "\*" "\*" " "   
## 11 ( 1 ) "\*" "\*" " "   
## 12 ( 1 ) "\*" "\*" " "   
## 13 ( 1 ) "\*" "\*" " "   
## 14 ( 1 ) "\*" "\*" " "   
## 15 ( 1 ) "\*" "\*" " "   
## 16 ( 1 ) "\*" "\*" " "   
## 17 ( 1 ) "\*" "\*" " "   
## 18 ( 1 ) "\*" "\*" " "   
## 19 ( 1 ) "\*" "\*" " "   
## 20 ( 1 ) "\*" "\*" " "   
## 21 ( 1 ) "\*" "\*" " "   
## 22 ( 1 ) "\*" "\*" "\*"   
## 23 ( 1 ) "\*" "\*" "\*"   
## 24 ( 1 ) "\*" "\*" "\*"   
## 25 ( 1 ) "\*" "\*" "\*"   
## 26 ( 1 ) "\*" "\*" "\*"   
## 27 ( 1 ) "\*" "\*" "\*"   
## 28 ( 1 ) "\*" "\*" "\*"   
## last\_fico\_range\_high acc\_now\_delinq inq\_last\_12m num\_tl\_30dpd  
## 1 ( 1 ) "\*" " " " " " "   
## 2 ( 1 ) "\*" " " " " " "   
## 3 ( 1 ) " " " " " " " "   
## 4 ( 1 ) "\*" " " " " " "   
## 5 ( 1 ) "\*" " " " " " "   
## 6 ( 1 ) "\*" " " " " " "   
## 7 ( 1 ) "\*" " " " " " "   
## 8 ( 1 ) "\*" " " " " " "   
## 9 ( 1 ) "\*" " " " " " "   
## 10 ( 1 ) "\*" " " " " " "   
## 11 ( 1 ) "\*" " " " " " "   
## 12 ( 1 ) "\*" " " " " " "   
## 13 ( 1 ) "\*" " " " " " "   
## 14 ( 1 ) "\*" " " " " " "   
## 15 ( 1 ) "\*" " " " " " "   
## 16 ( 1 ) "\*" " " " " " "   
## 17 ( 1 ) "\*" " " " " " "   
## 18 ( 1 ) "\*" " " " " "\*"   
## 19 ( 1 ) "\*" " " " " "\*"   
## 20 ( 1 ) "\*" " " " " "\*"   
## 21 ( 1 ) "\*" " " " " "\*"   
## 22 ( 1 ) "\*" " " " " "\*"   
## 23 ( 1 ) "\*" " " " " "\*"   
## 24 ( 1 ) "\*" " " " " "\*"   
## 25 ( 1 ) "\*" " " " " "\*"   
## 26 ( 1 ) "\*" " " " " "\*"   
## 27 ( 1 ) "\*" "\*" " " "\*"   
## 28 ( 1 ) "\*" "\*" "\*" "\*"   
## num\_tl\_90g\_dpd\_24m tot\_hi\_cred\_lim num\_rev\_accts  
## 1 ( 1 ) " " " " " "   
## 2 ( 1 ) " " " " " "   
## 3 ( 1 ) " " " " " "   
## 4 ( 1 ) " " " " " "   
## 5 ( 1 ) " " " " " "   
## 6 ( 1 ) " " " " " "   
## 7 ( 1 ) " " " " " "   
## 8 ( 1 ) " " " " " "   
## 9 ( 1 ) " " " " " "   
## 10 ( 1 ) " " " " " "   
## 11 ( 1 ) " " " " " "   
## 12 ( 1 ) " " " " " "   
## 13 ( 1 ) " " " " "\*"   
## 14 ( 1 ) " " " " "\*"   
## 15 ( 1 ) " " " " "\*"   
## 16 ( 1 ) " " "\*" "\*"   
## 17 ( 1 ) " " "\*" "\*"   
## 18 ( 1 ) " " "\*" "\*"   
## 19 ( 1 ) " " "\*" "\*"   
## 20 ( 1 ) " " "\*" "\*"   
## 21 ( 1 ) " " "\*" "\*"   
## 22 ( 1 ) " " "\*" "\*"   
## 23 ( 1 ) " " "\*" "\*"   
## 24 ( 1 ) " " "\*" "\*"   
## 25 ( 1 ) "\*" "\*" "\*"   
## 26 ( 1 ) "\*" "\*" "\*"   
## 27 ( 1 ) "\*" "\*" "\*"   
## 28 ( 1 ) "\*" "\*" "\*"

regular\_subset\_full\_summary$rsq

## [1] 0.1649569 0.2113047 0.2800062 0.3721668 0.3876568 0.4024289 0.4136643  
## [8] 0.4145148 0.4150604 0.4153697 0.4154948 0.4156485 0.4157870 0.4159196  
## [15] 0.4160709 0.4161246 0.4161593 0.4161740 0.4161873 0.4161973 0.4162080  
## [22] 0.4162178 0.4162256 0.4162328 0.4162372 0.4162396 0.4162410 0.4162413

par(mfrow =c(2,2))  
plot(regular\_subset\_full\_summary$rss ,xlab=" Number of Variables", ylab=" RSS", type="l")  
plot(regular\_subset\_full\_summary$adjr2 ,xlab =" Number of Variables", ylab=" Adjusted RSq",type="l")  
points (which.max(regular\_subset\_full\_summary$adjr2), regular\_subset\_full\_summary$adjr2[which.max(regular\_subset\_full\_summary$adjr2)], col ="red",cex =2, pch =20)  
plot(regular\_subset\_full\_summary$cp ,xlab =" Number of Variables ",ylab="Cp", type="l")  
points (which.min (regular\_subset\_full\_summary$cp ), regular\_subset\_full\_summary$cp [which.min (regular\_subset\_full\_summary$cp )], col ="red",cex =2, pch =20)  
plot(regular\_subset\_full\_summary$bic ,xlab=" Number of Variables ",ylab="BIC", type="l")  
points(which.min (regular\_subset\_full\_summary$bic), regular\_subset\_full\_summary$bic[which.min(regular\_subset\_full\_summary$bic )], col ="red",cex =2, pch =20)



## Forward and backward subset selection test.

forward\_subset\_full <- regsubsets(is\_acct\_delinquent ~ ., ds\_lc, nvmax = 40, method="forward")  
summary(forward\_subset\_full)

## Subset selection object  
## Call: regsubsets.formula(is\_acct\_delinquent ~ ., ds\_lc, nvmax = 40,   
## method = "forward")  
## 28 Variables (and intercept)  
## Forced in Forced out  
## loan\_amnt FALSE FALSE  
## term FALSE FALSE  
## int\_rate FALSE FALSE  
## installment FALSE FALSE  
## home\_ownership FALSE FALSE  
## annual\_inc FALSE FALSE  
## dti FALSE FALSE  
## delinq\_2yrs FALSE FALSE  
## fico\_range\_high FALSE FALSE  
## inq\_last\_6mths FALSE FALSE  
## open\_acc FALSE FALSE  
## pub\_rec FALSE FALSE  
## revol\_bal FALSE FALSE  
## total\_acc FALSE FALSE  
## out\_prncp FALSE FALSE  
## out\_prncp\_inv FALSE FALSE  
## total\_pymnt FALSE FALSE  
## total\_rec\_int FALSE FALSE  
## total\_rec\_late\_fee FALSE FALSE  
## recoveries FALSE FALSE  
## collection\_recovery\_fee FALSE FALSE  
## last\_fico\_range\_high FALSE FALSE  
## acc\_now\_delinq FALSE FALSE  
## inq\_last\_12m FALSE FALSE  
## num\_tl\_30dpd FALSE FALSE  
## num\_tl\_90g\_dpd\_24m FALSE FALSE  
## tot\_hi\_cred\_lim FALSE FALSE  
## num\_rev\_accts FALSE FALSE  
## 1 subsets of each size up to 28  
## Selection Algorithm: forward  
## loan\_amnt term int\_rate installment home\_ownership annual\_inc dti  
## 1 ( 1 ) " " " " " " " " " " " " " "  
## 2 ( 1 ) " " " " " " " " " " " " " "  
## 3 ( 1 ) " " " " " " " " " " " " " "  
## 4 ( 1 ) " " " " " " " " " " " " " "  
## 5 ( 1 ) "\*" " " " " " " " " " " " "  
## 6 ( 1 ) "\*" " " " " " " " " " " " "  
## 7 ( 1 ) "\*" " " " " " " " " " " " "  
## 8 ( 1 ) "\*" " " " " " " " " " " " "  
## 9 ( 1 ) "\*" " " "\*" " " " " " " " "  
## 10 ( 1 ) "\*" " " "\*" " " " " "\*" " "  
## 11 ( 1 ) "\*" " " "\*" " " " " "\*" " "  
## 12 ( 1 ) "\*" " " "\*" "\*" " " "\*" " "  
## 13 ( 1 ) "\*" "\*" "\*" "\*" " " "\*" " "  
## 14 ( 1 ) "\*" "\*" "\*" "\*" " " "\*" " "  
## 15 ( 1 ) "\*" "\*" "\*" "\*" " " "\*" " "  
## 16 ( 1 ) "\*" "\*" "\*" "\*" " " "\*" " "  
## 17 ( 1 ) "\*" "\*" "\*" "\*" " " "\*" "\*"  
## 18 ( 1 ) "\*" "\*" "\*" "\*" " " "\*" "\*"  
## 19 ( 1 ) "\*" "\*" "\*" "\*" " " "\*" "\*"  
## 20 ( 1 ) "\*" "\*" "\*" "\*" " " "\*" "\*"  
## 21 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*"  
## 22 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*"  
## 23 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*"  
## 24 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*"  
## 25 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*"  
## 26 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*"  
## 27 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*"  
## 28 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*"  
## delinq\_2yrs fico\_range\_high inq\_last\_6mths open\_acc pub\_rec revol\_bal  
## 1 ( 1 ) " " " " " " " " " " " "   
## 2 ( 1 ) " " "\*" " " " " " " " "   
## 3 ( 1 ) " " "\*" " " " " " " " "   
## 4 ( 1 ) " " "\*" " " " " " " " "   
## 5 ( 1 ) " " "\*" " " " " " " " "   
## 6 ( 1 ) " " "\*" " " " " " " " "   
## 7 ( 1 ) " " "\*" " " " " " " " "   
## 8 ( 1 ) " " "\*" " " " " " " " "   
## 9 ( 1 ) " " "\*" " " " " " " " "   
## 10 ( 1 ) " " "\*" " " " " " " " "   
## 11 ( 1 ) "\*" "\*" " " " " " " " "   
## 12 ( 1 ) "\*" "\*" " " " " " " " "   
## 13 ( 1 ) "\*" "\*" " " " " " " " "   
## 14 ( 1 ) "\*" "\*" " " " " " " " "   
## 15 ( 1 ) "\*" "\*" " " " " " " " "   
## 16 ( 1 ) "\*" "\*" " " "\*" " " " "   
## 17 ( 1 ) "\*" "\*" " " "\*" " " " "   
## 18 ( 1 ) "\*" "\*" " " "\*" " " " "   
## 19 ( 1 ) "\*" "\*" " " "\*" "\*" " "   
## 20 ( 1 ) "\*" "\*" " " "\*" "\*" " "   
## 21 ( 1 ) "\*" "\*" " " "\*" "\*" " "   
## 22 ( 1 ) "\*" "\*" " " "\*" "\*" " "   
## 23 ( 1 ) "\*" "\*" " " "\*" "\*" " "   
## 24 ( 1 ) "\*" "\*" " " "\*" "\*" "\*"   
## 25 ( 1 ) "\*" "\*" " " "\*" "\*" "\*"   
## 26 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*"   
## 27 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*"   
## 28 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*"   
## total\_acc out\_prncp out\_prncp\_inv total\_pymnt total\_rec\_int  
## 1 ( 1 ) " " " " " " " " " "   
## 2 ( 1 ) " " " " " " " " " "   
## 3 ( 1 ) " " " " " " " " " "   
## 4 ( 1 ) " " " " " " " " " "   
## 5 ( 1 ) " " " " " " " " " "   
## 6 ( 1 ) " " "\*" " " " " " "   
## 7 ( 1 ) " " "\*" " " "\*" " "   
## 8 ( 1 ) " " "\*" " " "\*" "\*"   
## 9 ( 1 ) " " "\*" " " "\*" "\*"   
## 10 ( 1 ) " " "\*" " " "\*" "\*"   
## 11 ( 1 ) " " "\*" " " "\*" "\*"   
## 12 ( 1 ) " " "\*" " " "\*" "\*"   
## 13 ( 1 ) " " "\*" " " "\*" "\*"   
## 14 ( 1 ) " " "\*" " " "\*" "\*"   
## 15 ( 1 ) " " "\*" " " "\*" "\*"   
## 16 ( 1 ) " " "\*" " " "\*" "\*"   
## 17 ( 1 ) " " "\*" " " "\*" "\*"   
## 18 ( 1 ) " " "\*" " " "\*" "\*"   
## 19 ( 1 ) " " "\*" " " "\*" "\*"   
## 20 ( 1 ) "\*" "\*" " " "\*" "\*"   
## 21 ( 1 ) "\*" "\*" " " "\*" "\*"   
## 22 ( 1 ) "\*" "\*" " " "\*" "\*"   
## 23 ( 1 ) "\*" "\*" "\*" "\*" "\*"   
## 24 ( 1 ) "\*" "\*" "\*" "\*" "\*"   
## 25 ( 1 ) "\*" "\*" "\*" "\*" "\*"   
## 26 ( 1 ) "\*" "\*" "\*" "\*" "\*"   
## 27 ( 1 ) "\*" "\*" "\*" "\*" "\*"   
## 28 ( 1 ) "\*" "\*" "\*" "\*" "\*"   
## total\_rec\_late\_fee recoveries collection\_recovery\_fee  
## 1 ( 1 ) " " " " " "   
## 2 ( 1 ) " " " " " "   
## 3 ( 1 ) " " "\*" " "   
## 4 ( 1 ) "\*" "\*" " "   
## 5 ( 1 ) "\*" "\*" " "   
## 6 ( 1 ) "\*" "\*" " "   
## 7 ( 1 ) "\*" "\*" " "   
## 8 ( 1 ) "\*" "\*" " "   
## 9 ( 1 ) "\*" "\*" " "   
## 10 ( 1 ) "\*" "\*" " "   
## 11 ( 1 ) "\*" "\*" " "   
## 12 ( 1 ) "\*" "\*" " "   
## 13 ( 1 ) "\*" "\*" " "   
## 14 ( 1 ) "\*" "\*" " "   
## 15 ( 1 ) "\*" "\*" " "   
## 16 ( 1 ) "\*" "\*" " "   
## 17 ( 1 ) "\*" "\*" " "   
## 18 ( 1 ) "\*" "\*" " "   
## 19 ( 1 ) "\*" "\*" " "   
## 20 ( 1 ) "\*" "\*" " "   
## 21 ( 1 ) "\*" "\*" " "   
## 22 ( 1 ) "\*" "\*" "\*"   
## 23 ( 1 ) "\*" "\*" "\*"   
## 24 ( 1 ) "\*" "\*" "\*"   
## 25 ( 1 ) "\*" "\*" "\*"   
## 26 ( 1 ) "\*" "\*" "\*"   
## 27 ( 1 ) "\*" "\*" "\*"   
## 28 ( 1 ) "\*" "\*" "\*"   
## last\_fico\_range\_high acc\_now\_delinq inq\_last\_12m num\_tl\_30dpd  
## 1 ( 1 ) "\*" " " " " " "   
## 2 ( 1 ) "\*" " " " " " "   
## 3 ( 1 ) "\*" " " " " " "   
## 4 ( 1 ) "\*" " " " " " "   
## 5 ( 1 ) "\*" " " " " " "   
## 6 ( 1 ) "\*" " " " " " "   
## 7 ( 1 ) "\*" " " " " " "   
## 8 ( 1 ) "\*" " " " " " "   
## 9 ( 1 ) "\*" " " " " " "   
## 10 ( 1 ) "\*" " " " " " "   
## 11 ( 1 ) "\*" " " " " " "   
## 12 ( 1 ) "\*" " " " " " "   
## 13 ( 1 ) "\*" " " " " " "   
## 14 ( 1 ) "\*" " " " " " "   
## 15 ( 1 ) "\*" " " " " " "   
## 16 ( 1 ) "\*" " " " " " "   
## 17 ( 1 ) "\*" " " " " " "   
## 18 ( 1 ) "\*" " " " " "\*"   
## 19 ( 1 ) "\*" " " " " "\*"   
## 20 ( 1 ) "\*" " " " " "\*"   
## 21 ( 1 ) "\*" " " " " "\*"   
## 22 ( 1 ) "\*" " " " " "\*"   
## 23 ( 1 ) "\*" " " " " "\*"   
## 24 ( 1 ) "\*" " " " " "\*"   
## 25 ( 1 ) "\*" " " " " "\*"   
## 26 ( 1 ) "\*" " " " " "\*"   
## 27 ( 1 ) "\*" "\*" " " "\*"   
## 28 ( 1 ) "\*" "\*" "\*" "\*"   
## num\_tl\_90g\_dpd\_24m tot\_hi\_cred\_lim num\_rev\_accts  
## 1 ( 1 ) " " " " " "   
## 2 ( 1 ) " " " " " "   
## 3 ( 1 ) " " " " " "   
## 4 ( 1 ) " " " " " "   
## 5 ( 1 ) " " " " " "   
## 6 ( 1 ) " " " " " "   
## 7 ( 1 ) " " " " " "   
## 8 ( 1 ) " " " " " "   
## 9 ( 1 ) " " " " " "   
## 10 ( 1 ) " " " " " "   
## 11 ( 1 ) " " " " " "   
## 12 ( 1 ) " " " " " "   
## 13 ( 1 ) " " " " " "   
## 14 ( 1 ) " " "\*" " "   
## 15 ( 1 ) " " "\*" "\*"   
## 16 ( 1 ) " " "\*" "\*"   
## 17 ( 1 ) " " "\*" "\*"   
## 18 ( 1 ) " " "\*" "\*"   
## 19 ( 1 ) " " "\*" "\*"   
## 20 ( 1 ) " " "\*" "\*"   
## 21 ( 1 ) " " "\*" "\*"   
## 22 ( 1 ) " " "\*" "\*"   
## 23 ( 1 ) " " "\*" "\*"   
## 24 ( 1 ) " " "\*" "\*"   
## 25 ( 1 ) "\*" "\*" "\*"   
## 26 ( 1 ) "\*" "\*" "\*"   
## 27 ( 1 ) "\*" "\*" "\*"   
## 28 ( 1 ) "\*" "\*" "\*"

backward\_subset\_full <- regsubsets(is\_acct\_delinquent ~ ., ds\_lc, nvmax = 40, method="backward")  
summary(backward\_subset\_full)

## Subset selection object  
## Call: regsubsets.formula(is\_acct\_delinquent ~ ., ds\_lc, nvmax = 40,   
## method = "backward")  
## 28 Variables (and intercept)  
## Forced in Forced out  
## loan\_amnt FALSE FALSE  
## term FALSE FALSE  
## int\_rate FALSE FALSE  
## installment FALSE FALSE  
## home\_ownership FALSE FALSE  
## annual\_inc FALSE FALSE  
## dti FALSE FALSE  
## delinq\_2yrs FALSE FALSE  
## fico\_range\_high FALSE FALSE  
## inq\_last\_6mths FALSE FALSE  
## open\_acc FALSE FALSE  
## pub\_rec FALSE FALSE  
## revol\_bal FALSE FALSE  
## total\_acc FALSE FALSE  
## out\_prncp FALSE FALSE  
## out\_prncp\_inv FALSE FALSE  
## total\_pymnt FALSE FALSE  
## total\_rec\_int FALSE FALSE  
## total\_rec\_late\_fee FALSE FALSE  
## recoveries FALSE FALSE  
## collection\_recovery\_fee FALSE FALSE  
## last\_fico\_range\_high FALSE FALSE  
## acc\_now\_delinq FALSE FALSE  
## inq\_last\_12m FALSE FALSE  
## num\_tl\_30dpd FALSE FALSE  
## num\_tl\_90g\_dpd\_24m FALSE FALSE  
## tot\_hi\_cred\_lim FALSE FALSE  
## num\_rev\_accts FALSE FALSE  
## 1 subsets of each size up to 28  
## Selection Algorithm: backward  
## loan\_amnt term int\_rate installment home\_ownership annual\_inc dti  
## 1 ( 1 ) " " " " " " " " " " " " " "  
## 2 ( 1 ) "\*" " " " " " " " " " " " "  
## 3 ( 1 ) "\*" " " " " " " " " " " " "  
## 4 ( 1 ) "\*" " " " " " " " " " " " "  
## 5 ( 1 ) "\*" " " " " " " " " " " " "  
## 6 ( 1 ) "\*" " " " " " " " " " " " "  
## 7 ( 1 ) "\*" " " " " " " " " " " " "  
## 8 ( 1 ) "\*" " " "\*" " " " " " " " "  
## 9 ( 1 ) "\*" " " "\*" " " " " " " " "  
## 10 ( 1 ) "\*" " " "\*" " " " " "\*" " "  
## 11 ( 1 ) "\*" " " "\*" " " " " "\*" " "  
## 12 ( 1 ) "\*" " " "\*" " " " " "\*" " "  
## 13 ( 1 ) "\*" " " "\*" "\*" " " "\*" " "  
## 14 ( 1 ) "\*" "\*" "\*" "\*" " " "\*" " "  
## 15 ( 1 ) "\*" "\*" "\*" "\*" " " "\*" " "  
## 16 ( 1 ) "\*" "\*" "\*" "\*" " " "\*" " "  
## 17 ( 1 ) "\*" "\*" "\*" "\*" " " "\*" "\*"  
## 18 ( 1 ) "\*" "\*" "\*" "\*" " " "\*" "\*"  
## 19 ( 1 ) "\*" "\*" "\*" "\*" " " "\*" "\*"  
## 20 ( 1 ) "\*" "\*" "\*" "\*" " " "\*" "\*"  
## 21 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*"  
## 22 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*"  
## 23 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*"  
## 24 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*"  
## 25 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*"  
## 26 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*"  
## 27 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*"  
## 28 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*"  
## delinq\_2yrs fico\_range\_high inq\_last\_6mths open\_acc pub\_rec revol\_bal  
## 1 ( 1 ) " " " " " " " " " " " "   
## 2 ( 1 ) " " " " " " " " " " " "   
## 3 ( 1 ) " " " " " " " " " " " "   
## 4 ( 1 ) " " " " " " " " " " " "   
## 5 ( 1 ) " " " " " " " " " " " "   
## 6 ( 1 ) " " "\*" " " " " " " " "   
## 7 ( 1 ) " " "\*" " " " " " " " "   
## 8 ( 1 ) " " "\*" " " " " " " " "   
## 9 ( 1 ) " " "\*" " " " " " " " "   
## 10 ( 1 ) " " "\*" " " " " " " " "   
## 11 ( 1 ) " " "\*" " " "\*" " " " "   
## 12 ( 1 ) " " "\*" " " "\*" " " " "   
## 13 ( 1 ) " " "\*" " " "\*" " " " "   
## 14 ( 1 ) " " "\*" " " "\*" " " " "   
## 15 ( 1 ) "\*" "\*" " " "\*" " " " "   
## 16 ( 1 ) "\*" "\*" " " "\*" " " " "   
## 17 ( 1 ) "\*" "\*" " " "\*" " " " "   
## 18 ( 1 ) "\*" "\*" " " "\*" " " " "   
## 19 ( 1 ) "\*" "\*" " " "\*" "\*" " "   
## 20 ( 1 ) "\*" "\*" " " "\*" "\*" " "   
## 21 ( 1 ) "\*" "\*" " " "\*" "\*" " "   
## 22 ( 1 ) "\*" "\*" " " "\*" "\*" " "   
## 23 ( 1 ) "\*" "\*" " " "\*" "\*" " "   
## 24 ( 1 ) "\*" "\*" " " "\*" "\*" "\*"   
## 25 ( 1 ) "\*" "\*" " " "\*" "\*" "\*"   
## 26 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*"   
## 27 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*"   
## 28 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*"   
## total\_acc out\_prncp out\_prncp\_inv total\_pymnt total\_rec\_int  
## 1 ( 1 ) " " "\*" " " " " " "   
## 2 ( 1 ) " " "\*" " " " " " "   
## 3 ( 1 ) " " "\*" " " "\*" " "   
## 4 ( 1 ) " " "\*" " " "\*" " "   
## 5 ( 1 ) " " "\*" " " "\*" "\*"   
## 6 ( 1 ) " " "\*" " " "\*" "\*"   
## 7 ( 1 ) " " "\*" " " "\*" "\*"   
## 8 ( 1 ) " " "\*" " " "\*" "\*"   
## 9 ( 1 ) " " "\*" " " "\*" "\*"   
## 10 ( 1 ) " " "\*" " " "\*" "\*"   
## 11 ( 1 ) " " "\*" " " "\*" "\*"   
## 12 ( 1 ) " " "\*" " " "\*" "\*"   
## 13 ( 1 ) " " "\*" " " "\*" "\*"   
## 14 ( 1 ) " " "\*" " " "\*" "\*"   
## 15 ( 1 ) " " "\*" " " "\*" "\*"   
## 16 ( 1 ) " " "\*" " " "\*" "\*"   
## 17 ( 1 ) " " "\*" " " "\*" "\*"   
## 18 ( 1 ) " " "\*" " " "\*" "\*"   
## 19 ( 1 ) " " "\*" " " "\*" "\*"   
## 20 ( 1 ) "\*" "\*" " " "\*" "\*"   
## 21 ( 1 ) "\*" "\*" " " "\*" "\*"   
## 22 ( 1 ) "\*" "\*" " " "\*" "\*"   
## 23 ( 1 ) "\*" "\*" "\*" "\*" "\*"   
## 24 ( 1 ) "\*" "\*" "\*" "\*" "\*"   
## 25 ( 1 ) "\*" "\*" "\*" "\*" "\*"   
## 26 ( 1 ) "\*" "\*" "\*" "\*" "\*"   
## 27 ( 1 ) "\*" "\*" "\*" "\*" "\*"   
## 28 ( 1 ) "\*" "\*" "\*" "\*" "\*"   
## total\_rec\_late\_fee recoveries collection\_recovery\_fee  
## 1 ( 1 ) " " " " " "   
## 2 ( 1 ) " " " " " "   
## 3 ( 1 ) " " " " " "   
## 4 ( 1 ) " " " " " "   
## 5 ( 1 ) " " " " " "   
## 6 ( 1 ) " " " " " "   
## 7 ( 1 ) "\*" " " " "   
## 8 ( 1 ) "\*" " " " "   
## 9 ( 1 ) "\*" "\*" " "   
## 10 ( 1 ) "\*" "\*" " "   
## 11 ( 1 ) "\*" "\*" " "   
## 12 ( 1 ) "\*" "\*" " "   
## 13 ( 1 ) "\*" "\*" " "   
## 14 ( 1 ) "\*" "\*" " "   
## 15 ( 1 ) "\*" "\*" " "   
## 16 ( 1 ) "\*" "\*" " "   
## 17 ( 1 ) "\*" "\*" " "   
## 18 ( 1 ) "\*" "\*" " "   
## 19 ( 1 ) "\*" "\*" " "   
## 20 ( 1 ) "\*" "\*" " "   
## 21 ( 1 ) "\*" "\*" " "   
## 22 ( 1 ) "\*" "\*" "\*"   
## 23 ( 1 ) "\*" "\*" "\*"   
## 24 ( 1 ) "\*" "\*" "\*"   
## 25 ( 1 ) "\*" "\*" "\*"   
## 26 ( 1 ) "\*" "\*" "\*"   
## 27 ( 1 ) "\*" "\*" "\*"   
## 28 ( 1 ) "\*" "\*" "\*"   
## last\_fico\_range\_high acc\_now\_delinq inq\_last\_12m num\_tl\_30dpd  
## 1 ( 1 ) " " " " " " " "   
## 2 ( 1 ) " " " " " " " "   
## 3 ( 1 ) " " " " " " " "   
## 4 ( 1 ) "\*" " " " " " "   
## 5 ( 1 ) "\*" " " " " " "   
## 6 ( 1 ) "\*" " " " " " "   
## 7 ( 1 ) "\*" " " " " " "   
## 8 ( 1 ) "\*" " " " " " "   
## 9 ( 1 ) "\*" " " " " " "   
## 10 ( 1 ) "\*" " " " " " "   
## 11 ( 1 ) "\*" " " " " " "   
## 12 ( 1 ) "\*" " " " " " "   
## 13 ( 1 ) "\*" " " " " " "   
## 14 ( 1 ) "\*" " " " " " "   
## 15 ( 1 ) "\*" " " " " " "   
## 16 ( 1 ) "\*" " " " " " "   
## 17 ( 1 ) "\*" " " " " " "   
## 18 ( 1 ) "\*" " " " " "\*"   
## 19 ( 1 ) "\*" " " " " "\*"   
## 20 ( 1 ) "\*" " " " " "\*"   
## 21 ( 1 ) "\*" " " " " "\*"   
## 22 ( 1 ) "\*" " " " " "\*"   
## 23 ( 1 ) "\*" " " " " "\*"   
## 24 ( 1 ) "\*" " " " " "\*"   
## 25 ( 1 ) "\*" " " " " "\*"   
## 26 ( 1 ) "\*" " " " " "\*"   
## 27 ( 1 ) "\*" "\*" " " "\*"   
## 28 ( 1 ) "\*" "\*" "\*" "\*"   
## num\_tl\_90g\_dpd\_24m tot\_hi\_cred\_lim num\_rev\_accts  
## 1 ( 1 ) " " " " " "   
## 2 ( 1 ) " " " " " "   
## 3 ( 1 ) " " " " " "   
## 4 ( 1 ) " " " " " "   
## 5 ( 1 ) " " " " " "   
## 6 ( 1 ) " " " " " "   
## 7 ( 1 ) " " " " " "   
## 8 ( 1 ) " " " " " "   
## 9 ( 1 ) " " " " " "   
## 10 ( 1 ) " " " " " "   
## 11 ( 1 ) " " " " " "   
## 12 ( 1 ) " " " " "\*"   
## 13 ( 1 ) " " " " "\*"   
## 14 ( 1 ) " " " " "\*"   
## 15 ( 1 ) " " " " "\*"   
## 16 ( 1 ) " " "\*" "\*"   
## 17 ( 1 ) " " "\*" "\*"   
## 18 ( 1 ) " " "\*" "\*"   
## 19 ( 1 ) " " "\*" "\*"   
## 20 ( 1 ) " " "\*" "\*"   
## 21 ( 1 ) " " "\*" "\*"   
## 22 ( 1 ) " " "\*" "\*"   
## 23 ( 1 ) " " "\*" "\*"   
## 24 ( 1 ) " " "\*" "\*"   
## 25 ( 1 ) "\*" "\*" "\*"   
## 26 ( 1 ) "\*" "\*" "\*"   
## 27 ( 1 ) "\*" "\*" "\*"   
## 28 ( 1 ) "\*" "\*" "\*"

Based on the best subset selection dropping all the below variables

ds\_lc = subset(ds\_lc, select = -c(out\_prncp\_inv,num\_tl\_30dpd,num\_tl\_90g\_dpd\_24m,num\_rev\_accts,acc\_now\_delinq, recoveries, collection\_recovery\_fee))

Check data types for all variables is int or numeric

str(ds\_lc)

## Classes 'data.table' and 'data.frame': 494110 obs. of 22 variables:  
## $ loan\_amnt : int 5000 15000 11200 25000 3000 17000 20000 19200 6500 10000 ...  
## $ term : int 1 1 2 2 1 2 1 2 1 2 ...  
## $ int\_rate : num 20.39 9.92 30.79 21.85 7.34 ...  
## $ installment : num 186.8 483.4 367.8 688.4 93.1 ...  
## $ home\_ownership : int 4 3 4 2 4 4 2 4 2 2 ...  
## $ annual\_inc : num 50000 196000 44000 65000 52000 52000 19000 36500 50000 80000 ...  
## $ dti : num 21.8 18.29 43.97 12.89 0.58 ...  
## $ delinq\_2yrs : int 1 0 1 1 0 0 0 3 0 0 ...  
## $ fico\_range\_high : int 669 704 669 669 764 674 799 679 709 664 ...  
## $ inq\_last\_6mths : int 0 0 2 1 0 0 1 0 0 1 ...  
## $ open\_acc : int 5 19 8 7 7 9 6 14 7 17 ...  
## $ pub\_rec : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ revol\_bal : int 116 24243 1526 8657 141 15363 17144 13295 7871 10710 ...  
## $ total\_acc : int 18 53 14 16 30 14 9 18 16 23 ...  
## $ out\_prncp : num 3780 10878 10194 22189 0 ...  
## $ total\_pymnt : num 2044 5301 4008 7511 3012 ...  
## $ total\_rec\_int : num 824 1179.9 3001.4 4699.9 11.6 ...  
## $ total\_rec\_late\_fee : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ last\_fico\_range\_high: int 609 694 629 669 764 634 799 559 714 634 ...  
## $ inq\_last\_12m : int 5 7 10 3 2 2 1 0 0 4 ...  
## $ tot\_hi\_cred\_lim : num 33430 605228 80367 101234 191216 ...  
## $ is\_acct\_delinquent : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...  
## - attr(\*, ".internal.selfref")=<externalptr>

Summary for all variables in the dataset along with NA check.

summary(ds\_lc)

## loan\_amnt term int\_rate installment   
## Min. : 1000 Min. :1.000 Min. : 5.31 Min. : 29.76   
## 1st Qu.: 8000 1st Qu.:1.000 1st Qu.: 8.46 1st Qu.: 254.50   
## Median :14000 Median :1.000 Median :11.80 Median : 386.76   
## Mean :16017 Mean :1.304 Mean :12.73 Mean : 466.38   
## 3rd Qu.:22000 3rd Qu.:2.000 3rd Qu.:16.01 3rd Qu.: 628.67   
## Max. :40000 Max. :2.000 Max. :30.99 Max. :1670.15   
## home\_ownership annual\_inc dti delinq\_2yrs   
## Min. :1.00 Min. : 0 Min. : 0.00 Min. : 0.0000   
## 1st Qu.:2.00 1st Qu.: 46000 1st Qu.: 11.43 1st Qu.: 0.0000   
## Median :3.00 Median : 66000 Median : 17.71 Median : 0.0000   
## Mean :2.91 Mean : 80277 Mean : 19.67 Mean : 0.2295   
## 3rd Qu.:4.00 3rd Qu.: 96000 3rd Qu.: 25.03 3rd Qu.: 0.0000   
## Max. :4.00 Max. :9930475 Max. :999.00 Max. :58.0000   
## fico\_range\_high inq\_last\_6mths open\_acc pub\_rec   
## Min. :664.0 Min. :0.0000 Min. : 0.0 Min. : 0.0000   
## 1st Qu.:684.0 1st Qu.:0.0000 1st Qu.: 7.0 1st Qu.: 0.0000   
## Median :704.0 Median :0.0000 Median : 10.0 Median : 0.0000   
## Mean :710.4 Mean :0.4426 Mean : 11.5 Mean : 0.1346   
## 3rd Qu.:729.0 3rd Qu.:1.0000 3rd Qu.: 14.0 3rd Qu.: 0.0000   
## Max. :850.0 Max. :5.0000 Max. :101.0 Max. :52.0000   
## revol\_bal total\_acc out\_prncp total\_pymnt   
## Min. : 0 Min. : 2.00 Min. : 0 Min. : 0   
## 1st Qu.: 5306 1st Qu.: 14.00 1st Qu.: 4623 1st Qu.: 1757   
## Median : 10833 Median : 21.00 Median : 9736 Median : 3220   
## Mean : 16268 Mean : 22.64 Mean :12001 Mean : 4922   
## 3rd Qu.: 19865 3rd Qu.: 29.00 3rd Qu.:17788 3rd Qu.: 5902   
## Max. :2358150 Max. :160.00 Max. :40000 Max. :51653   
## total\_rec\_int total\_rec\_late\_fee last\_fico\_range\_high inq\_last\_12m   
## Min. : 0.0 Min. : 0.0000 Min. : 0.0 Min. : 0.000   
## 1st Qu.: 401.4 1st Qu.: 0.0000 1st Qu.:679.0 1st Qu.: 0.000   
## Median : 814.2 Median : 0.0000 Median :709.0 Median : 1.000   
## Mean : 1178.5 Mean : 0.5104 Mean :709.4 Mean : 1.938   
## 3rd Qu.: 1573.2 3rd Qu.: 0.0000 3rd Qu.:744.0 3rd Qu.: 3.000   
## Max. :12795.0 Max. :458.5700 Max. :850.0 Max. :67.000   
## tot\_hi\_cred\_lim is\_acct\_delinquent  
## Min. : 0 0:475138   
## 1st Qu.: 52075 1: 18972   
## Median : 116346   
## Mean : 184961   
## 3rd Qu.: 269331   
## Max. :9999999

Peak at initial set of records.

head(ds\_lc)

## loan\_amnt term int\_rate installment home\_ownership annual\_inc dti  
## 1: 5000 1 20.39 186.82 4 50000 21.80  
## 2: 15000 1 9.92 483.45 3 196000 18.29  
## 3: 11200 2 30.79 367.82 4 44000 43.97  
## 4: 25000 2 21.85 688.35 2 65000 12.89  
## 5: 3000 1 7.34 93.10 4 52000 0.58  
## 6: 17000 2 20.39 454.10 4 52000 15.65  
## delinq\_2yrs fico\_range\_high inq\_last\_6mths open\_acc pub\_rec revol\_bal  
## 1: 1 669 0 5 0 116  
## 2: 0 704 0 19 0 24243  
## 3: 1 669 2 8 0 1526  
## 4: 1 669 1 7 0 8657  
## 5: 0 764 0 7 0 141  
## 6: 0 674 0 9 0 15363  
## total\_acc out\_prncp total\_pymnt total\_rec\_int total\_rec\_late\_fee  
## 1: 18 3780.31 2043.690 824.00 0  
## 2: 53 10878.50 5301.420 1179.92 0  
## 3: 14 10193.73 4007.700 3001.43 0  
## 4: 16 22188.73 7511.160 4699.89 0  
## 5: 30 0.00 3011.577 11.58 0  
## 6: 14 15019.77 4956.590 2976.36 0  
## last\_fico\_range\_high inq\_last\_12m tot\_hi\_cred\_lim is\_acct\_delinquent  
## 1: 609 5 33430 0  
## 2: 694 7 605228 0  
## 3: 629 10 80367 0  
## 4: 669 3 101234 0  
## 5: 764 2 191216 0  
## 6: 634 2 32500 0

## Validation set

Identify what’s right the subset distribution for training and test dataset

sample\_set<-sample(nrow(ds\_lc),nrow(ds\_lc)\*0.7)  
training\_dataset <- ds\_lc[sample\_set,]  
test\_dataset <- ds\_lc[-sample\_set,]  
regression\_model <- glm(is\_acct\_delinquent ~   
 loan\_amnt +  
 term +  
 int\_rate +  
 installment +  
 home\_ownership +  
 annual\_inc +  
 dti +  
 delinq\_2yrs +  
 fico\_range\_high +  
 inq\_last\_6mths +  
 open\_acc +  
 pub\_rec +  
 revol\_bal +  
 total\_acc +  
 out\_prncp +  
 total\_pymnt +  
 total\_rec\_int +  
 total\_rec\_late\_fee +  
 last\_fico\_range\_high +  
 inq\_last\_12m +  
 tot\_hi\_cred\_lim,   
 data = training\_dataset, family = "binomial")

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

probablity<-predict(regression\_model,test\_dataset,type="response")  
prediction<-ifelse(probablity > 0.5, 1, 0)  
validation\_error\_rate1 <- mean(prediction != test\_dataset$is\_acct\_delinquent)  
validation\_error\_rate1

## [1] 0.01904434

sample\_set<-sample(nrow(ds\_lc),nrow(ds\_lc)\*0.75)  
training\_dataset <- ds\_lc[sample\_set,]  
test\_dataset <- ds\_lc[-sample\_set,]  
regression\_model <- glm(is\_acct\_delinquent ~   
 loan\_amnt +  
 term +  
 int\_rate +  
 installment +  
 home\_ownership +  
 annual\_inc +  
 dti +  
 delinq\_2yrs +  
 fico\_range\_high +  
 inq\_last\_6mths +  
 open\_acc +  
 pub\_rec +  
 revol\_bal +  
 total\_acc +  
 out\_prncp +  
 total\_pymnt +  
 total\_rec\_int +  
 total\_rec\_late\_fee +  
 last\_fico\_range\_high +  
 inq\_last\_12m +  
 tot\_hi\_cred\_lim,   
 data = training\_dataset, family = "binomial")

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

probablity<-predict(regression\_model,test\_dataset,type="response")  
prediction<-ifelse(probablity > 0.5, 1, 0)  
validation\_error\_rate2 <- mean(prediction != test\_dataset$is\_acct\_delinquent)  
validation\_error\_rate2

## [1] 0.01921022

sample\_set<-sample(nrow(ds\_lc),nrow(ds\_lc)\*0.8)  
training\_dataset <- ds\_lc[sample\_set,]  
test\_dataset <- ds\_lc[-sample\_set,]  
regression\_model <- glm(is\_acct\_delinquent ~   
 loan\_amnt +  
 term +  
 int\_rate +  
 installment +  
 home\_ownership +  
 annual\_inc +  
 dti +  
 delinq\_2yrs +  
 fico\_range\_high +  
 inq\_last\_6mths +  
 open\_acc +  
 pub\_rec +  
 revol\_bal +  
 total\_acc +  
 out\_prncp +  
 total\_pymnt +  
 total\_rec\_int +  
 total\_rec\_late\_fee +  
 last\_fico\_range\_high +  
 inq\_last\_12m +  
 tot\_hi\_cred\_lim,   
 data = training\_dataset, family = "binomial")

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

probablity<-predict(regression\_model,test\_dataset,type="response")  
prediction<-ifelse(probablity > 0.5, 1, 0)  
validation\_error\_rate3 <- mean(prediction != test\_dataset$is\_acct\_delinquent)  
validation\_error\_rate3

## [1] 0.01868005

Following is the data split in training and test data ration along with validation error rate/misclassification

1. Tranining : Test -> 70 : 30, validation error rate 0.0190443. or 1.9044342%   
2. Tranining : Test -> 75 : 25, validation error rate 0.0192102. or 1.921022%   
3. Tranining : Test -> 80 : 20, validation error rate 0.0186801. or 1.8680051%

Validation error rate is different for different sample splits of taining and test dataset. This indicates that validation error rate varies by which observations are in the training/validation sets.

Partition the dataset based on results of above validation set approach. For all the 3 validation set validation error rate was marginally different, so going with 0.75 distribution.

data\_partition <- createDataPartition(y = ds\_lc$is\_acct\_delinquent, p = 0.75, list = FALSE)  
train\_dataset <- ds\_lc[data\_partition,]  
test\_dataset <- ds\_lc[-data\_partition,]

Number of observations in trianing and test dataset

nrow(train\_dataset)

## [1] 370583

nrow(test\_dataset)

## [1] 123527

## Logistic Regression

logistic\_regression <- glm(is\_acct\_delinquent ~   
 loan\_amnt +  
 term +  
 int\_rate +  
 installment +  
 home\_ownership +  
 annual\_inc +  
 dti +  
 delinq\_2yrs +  
 fico\_range\_high +  
 inq\_last\_6mths +  
 open\_acc +  
 pub\_rec +  
 revol\_bal +  
 total\_acc +  
 out\_prncp +  
 total\_pymnt +  
 total\_rec\_int +  
 total\_rec\_late\_fee +  
 last\_fico\_range\_high +  
 inq\_last\_12m +  
 tot\_hi\_cred\_lim   
 ,data = train\_dataset, family = binomial)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

predictions <- predict(logistic\_regression, test\_dataset, type="response")  
predicted\_direction <- ifelse(predictions > 0.5, 1, 0)  
error\_rate\_lr <- mean(predicted\_direction != test\_dataset$is\_acct\_delinquent)  
error\_rate\_lr

## [1] 0.01939657

summary(logistic\_regression)

##   
## Call:  
## glm(formula = is\_acct\_delinquent ~ loan\_amnt + term + int\_rate +   
## installment + home\_ownership + annual\_inc + dti + delinq\_2yrs +   
## fico\_range\_high + inq\_last\_6mths + open\_acc + pub\_rec + revol\_bal +   
## total\_acc + out\_prncp + total\_pymnt + total\_rec\_int + total\_rec\_late\_fee +   
## last\_fico\_range\_high + inq\_last\_12m + tot\_hi\_cred\_lim, family = binomial,   
## data = train\_dataset)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -5.8434 -0.1700 -0.1096 -0.0653 7.4068   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 5.287e+00 3.745e-01 14.116 < 2e-16 \*\*\*  
## loan\_amnt 6.304e-03 6.369e-04 9.896 < 2e-16 \*\*\*  
## term 2.836e-01 6.199e-02 4.575 4.77e-06 \*\*\*  
## int\_rate 3.501e-02 3.954e-03 8.855 < 2e-16 \*\*\*  
## installment 3.063e-03 2.494e-04 12.285 < 2e-16 \*\*\*  
## home\_ownership 3.420e-02 1.662e-02 2.058 0.03956 \*   
## annual\_inc -2.711e-06 3.242e-07 -8.360 < 2e-16 \*\*\*  
## dti 8.080e-04 4.470e-04 1.808 0.07065 .   
## delinq\_2yrs -4.568e-02 1.575e-02 -2.900 0.00373 \*\*   
## fico\_range\_high 1.382e-02 4.979e-04 27.765 < 2e-16 \*\*\*  
## inq\_last\_6mths 3.712e-02 1.830e-02 2.028 0.04254 \*   
## open\_acc -2.368e-02 3.271e-03 -7.240 4.47e-13 \*\*\*  
## pub\_rec 1.998e-01 2.943e-02 6.787 1.14e-11 \*\*\*  
## revol\_bal -1.750e-06 8.207e-07 -2.132 0.03298 \*   
## total\_acc 1.756e-02 1.556e-03 11.284 < 2e-16 \*\*\*  
## out\_prncp -6.310e-03 6.370e-04 -9.906 < 2e-16 \*\*\*  
## total\_pymnt -6.523e-03 6.383e-04 -10.220 < 2e-16 \*\*\*  
## total\_rec\_int 6.286e-03 6.399e-04 9.823 < 2e-16 \*\*\*  
## total\_rec\_late\_fee 4.373e-02 1.208e-03 36.185 < 2e-16 \*\*\*  
## last\_fico\_range\_high -2.992e-02 2.559e-04 -116.929 < 2e-16 \*\*\*  
## inq\_last\_12m 3.658e-03 5.701e-03 0.642 0.52109   
## tot\_hi\_cred\_lim -4.004e-08 1.053e-07 -0.380 0.70383   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 120672 on 370582 degrees of freedom  
## Residual deviance: 51060 on 370561 degrees of freedom  
## AIC: 51104  
##   
## Number of Fisher Scoring iterations: 14

Following variables are statistically significant loan\_amnt , term, int\_rate, installment, annual\_inc, fico\_range\_high, open\_acc, pub\_rec, total\_acc, out\_prncp, total\_pymnt, total\_rec\_int, total\_rec\_late\_fee, last\_fico\_range\_high.

Following are less significant delinq\_2yrs, home\_ownership, revol\_bal, inq\_last\_6mths

## LDA - Linear Discriminant analysis

lda\_model <- lda(is\_acct\_delinquent ~   
 loan\_amnt +  
 term +  
 int\_rate +  
 installment +  
 home\_ownership +  
 annual\_inc +  
 dti +  
 delinq\_2yrs +  
 fico\_range\_high +  
 inq\_last\_6mths +  
 open\_acc +  
 pub\_rec +  
 revol\_bal +  
 total\_acc +  
 out\_prncp +  
 total\_pymnt +  
 total\_rec\_int +  
 total\_rec\_late\_fee +  
 last\_fico\_range\_high +  
 inq\_last\_12m +  
 tot\_hi\_cred\_lim  
 ,data = train\_dataset)  
predictions <- predict(lda\_model, test\_dataset, type="response")  
confusion\_matrix <- table(predictions$class,   
 test\_dataset$is\_acct\_delinquent,   
 dnn = c("Predicted Status", "Observed Status"))  
confusion\_matrix

## Observed Status  
## Predicted Status 0 1  
## 0 118637 2533  
## 1 147 2210

error\_rate\_lda <- mean(predictions$class != test\_dataset$is\_acct\_delinquent)  
error\_rate\_lda

## [1] 0.02169566

lda\_model

## Call:  
## lda(is\_acct\_delinquent ~ loan\_amnt + term + int\_rate + installment +   
## home\_ownership + annual\_inc + dti + delinq\_2yrs + fico\_range\_high +   
## inq\_last\_6mths + open\_acc + pub\_rec + revol\_bal + total\_acc +   
## out\_prncp + total\_pymnt + total\_rec\_int + total\_rec\_late\_fee +   
## last\_fico\_range\_high + inq\_last\_12m + tot\_hi\_cred\_lim, data = train\_dataset)  
##   
## Prior probabilities of groups:  
## 0 1   
## 0.96160374 0.03839626   
##   
## Group means:  
## loan\_amnt term int\_rate installment home\_ownership annual\_inc dti  
## 0 15968.86 1.301843 12.59062 464.1567 2.904645 80360.59 19.64461  
## 1 17407.84 1.379928 16.08258 522.6086 3.057067 75473.63 20.29670  
## delinq\_2yrs fico\_range\_high inq\_last\_6mths open\_acc pub\_rec revol\_bal  
## 0 0.2271730 710.7269 0.4349692 11.51513 0.1341475 16317.1  
## 1 0.2764073 702.2274 0.6299810 10.95692 0.1605173 14352.2  
## total\_acc out\_prncp total\_pymnt total\_rec\_int total\_rec\_late\_fee  
## 0 22.65216 12166.254 4979.227 1176.297 0.3191504  
## 1 21.77841 8175.167 3325.927 1247.486 5.4479029  
## last\_fico\_range\_high inq\_last\_12m tot\_hi\_cred\_lim  
## 0 713.5734 1.912626 185747.3  
## 1 603.5727 2.503479 157559.4  
##   
## Coefficients of linear discriminants:  
## LD1  
## loan\_amnt 3.783664e-04  
## term 1.248940e-01  
## int\_rate 1.033248e-02  
## installment 6.765234e-04  
## home\_ownership 1.037697e-02  
## annual\_inc -3.593782e-07  
## dti 5.299934e-04  
## delinq\_2yrs -3.157674e-02  
## fico\_range\_high 8.773818e-03  
## inq\_last\_6mths 4.495268e-03  
## open\_acc -5.829063e-03  
## pub\_rec 2.258794e-02  
## revol\_bal -6.018502e-09  
## total\_acc 2.632893e-03  
## out\_prncp -3.832435e-04  
## total\_pymnt -3.976857e-04  
## total\_rec\_int 2.890081e-04  
## total\_rec\_late\_fee 3.770164e-02  
## last\_fico\_range\_high -1.325355e-02  
## inq\_last\_12m -1.273754e-03  
## tot\_hi\_cred\_lim -9.278842e-08

Based on the coefficient values, all the features are important predictors and significant.

## QDA - Quadratic Discriminant Analysis

qda\_model <- qda(is\_acct\_delinquent ~   
 loan\_amnt +  
 term +  
 int\_rate +  
 installment +  
 home\_ownership +  
 annual\_inc +  
 dti +  
 delinq\_2yrs +  
 fico\_range\_high +  
 inq\_last\_6mths +  
 open\_acc +  
 pub\_rec +  
 revol\_bal +  
 total\_acc +  
 out\_prncp +  
 last\_fico\_range\_high +  
 inq\_last\_12m +  
 tot\_hi\_cred\_lim  
 ,data = train\_dataset)  
predictions <- predict(qda\_model, test\_dataset, type="response")  
confusion\_matrix <- table(predictions$class,   
 test\_dataset$is\_acct\_delinquent,   
 dnn = c("Predicted Status", "Observed Status"))  
confusion\_matrix

## Observed Status  
## Predicted Status 0 1  
## 0 114275 2077  
## 1 4509 2666

error\_rate\_qda <- mean(predictions$class != test\_dataset$is\_acct\_delinquent)  
error\_rate\_qda

## [1] 0.05331628

qda\_model

## Call:  
## qda(is\_acct\_delinquent ~ loan\_amnt + term + int\_rate + installment +   
## home\_ownership + annual\_inc + dti + delinq\_2yrs + fico\_range\_high +   
## inq\_last\_6mths + open\_acc + pub\_rec + revol\_bal + total\_acc +   
## out\_prncp + last\_fico\_range\_high + inq\_last\_12m + tot\_hi\_cred\_lim,   
## data = train\_dataset)  
##   
## Prior probabilities of groups:  
## 0 1   
## 0.96160374 0.03839626   
##   
## Group means:  
## loan\_amnt term int\_rate installment home\_ownership annual\_inc dti  
## 0 15968.86 1.301843 12.59062 464.1567 2.904645 80360.59 19.64461  
## 1 17407.84 1.379928 16.08258 522.6086 3.057067 75473.63 20.29670  
## delinq\_2yrs fico\_range\_high inq\_last\_6mths open\_acc pub\_rec revol\_bal  
## 0 0.2271730 710.7269 0.4349692 11.51513 0.1341475 16317.1  
## 1 0.2764073 702.2274 0.6299810 10.95692 0.1605173 14352.2  
## total\_acc out\_prncp last\_fico\_range\_high inq\_last\_12m tot\_hi\_cred\_lim  
## 0 22.65216 12166.254 713.5734 1.912626 185747.3  
## 1 21.77841 8175.167 603.5727 2.503479 157559.4

## KNN

train\_dataset\_matrix <- as.matrix(head(train\_dataset, 80000))  
test\_dataset\_matrix <- as.matrix(head(test\_dataset, 80000))  
train\_dataset\_mini <- head(train\_dataset, 80000)  
predictions <- knn(train\_dataset\_matrix, test\_dataset\_matrix, train\_dataset\_mini$is\_acct\_delinquent, 5)  
confusion\_matrix <- table(predictions,   
 train\_dataset\_mini$is\_acct\_delinquent,   
 dnn = c("Predicted Status", "Observed Status"))  
confusion\_matrix

## Observed Status  
## Predicted Status 0 1  
## 0 73935 5264  
## 1 766 35

error\_rate\_knn\_5 <- mean(predictions != train\_dataset\_mini$is\_acct\_delinquent)  
error\_rate\_knn\_5

## [1] 0.075375

predictions <- knn(train\_dataset\_matrix, test\_dataset\_matrix, train\_dataset\_mini$is\_acct\_delinquent, 10)  
confusion\_matrix <- table(predictions,   
 train\_dataset\_mini$is\_acct\_delinquent,   
 dnn = c("Predicted Status", "Observed Status"))  
confusion\_matrix

## Observed Status  
## Predicted Status 0 1  
## 0 74166 5270  
## 1 535 29

error\_rate\_knn\_10 <- mean(predictions != train\_dataset\_mini$is\_acct\_delinquent)  
error\_rate\_knn\_10

## [1] 0.0725625

predictions <- knn(train\_dataset\_matrix, test\_dataset\_matrix, train\_dataset\_mini$is\_acct\_delinquent, 15)  
confusion\_matrix <- table(predictions,   
 train\_dataset\_mini$is\_acct\_delinquent,   
 dnn = c("Predicted Status", "Observed Status"))  
confusion\_matrix

## Observed Status  
## Predicted Status 0 1  
## 0 74258 5276  
## 1 443 23

error\_rate\_knn\_15 <- mean(predictions != train\_dataset\_mini$is\_acct\_delinquent)  
error\_rate\_knn\_15

## [1] 0.0714875

predictions <- knn(train\_dataset\_matrix, test\_dataset\_matrix, train\_dataset\_mini$is\_acct\_delinquent, 20)  
confusion\_matrix <- table(predictions,   
 train\_dataset\_mini$is\_acct\_delinquent,   
 dnn = c("Predicted Status", "Observed Status"))  
confusion\_matrix

## Observed Status  
## Predicted Status 0 1  
## 0 74313 5281  
## 1 388 18

error\_rate\_knn\_20 <- mean(predictions != train\_dataset\_mini$is\_acct\_delinquent)  
error\_rate\_knn\_20

## [1] 0.0708625

For KNN after 10 centers the error rate is marginally dropping, so will go with n = 10 as baseline for comparison with other models

## Tree

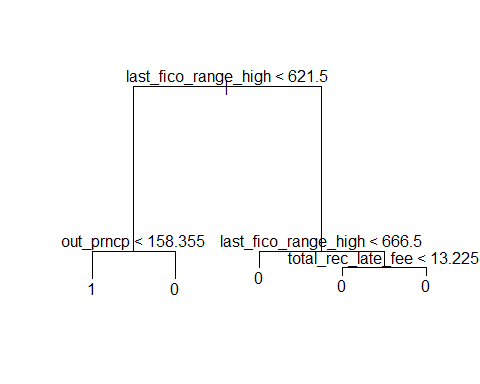
tree\_lc <- tree(is\_acct\_delinquent ~ .,train\_dataset)  
summary(tree\_lc)

##   
## Classification tree:  
## tree(formula = is\_acct\_delinquent ~ ., data = train\_dataset)  
## Variables actually used in tree construction:  
## [1] "last\_fico\_range\_high" "out\_prncp" "total\_rec\_late\_fee"   
## Number of terminal nodes: 5   
## Residual mean deviance: 0.1823 = 67560 / 370600   
## Misclassification error rate: 0.0248 = 9190 / 370583

tree\_lc

## node), split, n, deviance, yval, (yprob)  
## \* denotes terminal node  
##   
## 1) root 370583 120700 0 ( 0.961604 0.038396 )   
## 2) last\_fico\_range\_high < 621.5 16297 22540 1 ( 0.471191 0.528809 )   
## 4) out\_prncp < 158.355 5969 3266 1 ( 0.077902 0.922098 ) \*  
## 5) out\_prncp > 158.355 10328 12640 0 ( 0.698490 0.301510 ) \*  
## 3) last\_fico\_range\_high > 621.5 354286 57650 0 ( 0.984163 0.015837 )   
## 6) last\_fico\_range\_high < 666.5 40938 18900 0 ( 0.938590 0.061410 ) \*  
## 7) last\_fico\_range\_high > 666.5 313348 34760 0 ( 0.990116 0.009884 )   
## 14) total\_rec\_late\_fee < 13.225 311027 30450 0 ( 0.991509 0.008491 ) \*  
## 15) total\_rec\_late\_fee > 13.225 2321 2300 0 ( 0.803533 0.196467 ) \*

plot(tree\_lc)  
text(tree\_lc, pretty=0)



prediction <- predict(tree\_lc, test\_dataset, type="class")  
error\_rate\_tree <- mean(prediction != test\_dataset$is\_acct\_delinquent)  
error\_rate\_tree

## [1] 0.02543573

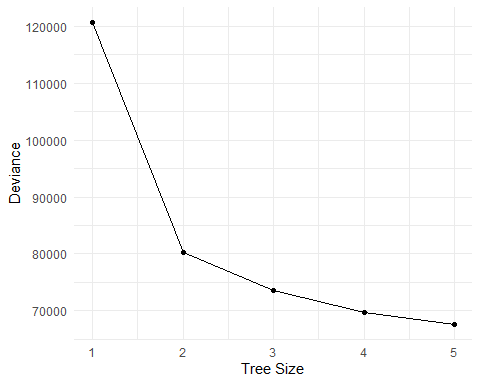
1. The prime indicator/factor for is loan account delinquent is last\_fico\_range\_high, out\_prncp and total\_rec\_late\_fee.
2. When last\_fico\_range\_high < 621.5 out\_prncp < 158.355 , then probablity of loan account becoming delinquent is high.
3. All other terminal nodes point towards not delinquent.

optimal\_tree <- cv.tree(tree\_lc, FUN = prune.tree)  
optimal\_tree

## $size  
## [1] 5 4 3 2 1  
##   
## $dev  
## [1] 67637.83 69646.40 73574.13 80195.54 120674.96  
##   
## $k  
## [1] -Inf 2012.395 3992.423 6627.537 40481.411  
##   
## $method  
## [1] "deviance"  
##   
## attr(,"class")  
## [1] "prune" "tree.sequence"

Optimal tree size is 5.

tree\_plot <- data.frame(x=optimal\_tree$size, y=optimal\_tree$dev)  
ggplot(tree\_plot, aes(x=x,y=y)) +   
 geom\_point() +   
 geom\_line() +   
 xlab("Tree Size") +   
 ylab("Deviance") +  
 theme\_minimal()



pruned\_tree <- prune.tree(tree\_lc, best = 5)  
optimal\_tree

## $size  
## [1] 5 4 3 2 1  
##   
## $dev  
## [1] 67637.83 69646.40 73574.13 80195.54 120674.96  
##   
## $k  
## [1] -Inf 2012.395 3992.423 6627.537 40481.411  
##   
## $method  
## [1] "deviance"  
##   
## attr(,"class")  
## [1] "prune" "tree.sequence"

summary(pruned\_tree)

##   
## Classification tree:  
## tree(formula = is\_acct\_delinquent ~ ., data = train\_dataset)  
## Variables actually used in tree construction:  
## [1] "last\_fico\_range\_high" "out\_prncp" "total\_rec\_late\_fee"   
## Number of terminal nodes: 5   
## Residual mean deviance: 0.1823 = 67560 / 370600   
## Misclassification error rate: 0.0248 = 9190 / 370583

summary(tree\_lc)

##   
## Classification tree:  
## tree(formula = is\_acct\_delinquent ~ ., data = train\_dataset)  
## Variables actually used in tree construction:  
## [1] "last\_fico\_range\_high" "out\_prncp" "total\_rec\_late\_fee"   
## Number of terminal nodes: 5   
## Residual mean deviance: 0.1823 = 67560 / 370600   
## Misclassification error rate: 0.0248 = 9190 / 370583

test\_error\_rate\_pruned <- mean(predict(pruned\_tree, test\_dataset, type = "class") != test\_dataset$is\_acct\_delinquent)  
test\_error\_rate\_pruned

## [1] 0.02543573

test\_error\_rate\_unpruned <- mean(prediction != test\_dataset$is\_acct\_delinquent)  
test\_error\_rate\_unpruned

## [1] 0.02543573

Test error rate of pruned tree is 0.0254357 and unpruned tree is 0.0254357.

Logistic regression error rate is 0.0193966. LDA error rate is 0.0216957. QDA error rate is 0.0533163. KNN error rate is 0.0725625. Tree error rate is 0.0254357. Unpruned tree error rate is 0.0254357.

# Predicting the Interest Rate

## Data Preparation, reset the training and test dataset, identify the interest mean for test dataset

Perform regular subset varibale selection test

ds\_lc <- ds\_lc\_after\_correlation  
regular\_subset\_full <- regsubsets(int\_rate ~ ., ds\_lc, nvmax = 40)  
regular\_subset\_full\_summary <- summary(regular\_subset\_full)  
regular\_subset\_full\_summary

## Subset selection object  
## Call: regsubsets.formula(int\_rate ~ ., ds\_lc, nvmax = 40)  
## 28 Variables (and intercept)  
## Forced in Forced out  
## loan\_amnt FALSE FALSE  
## term FALSE FALSE  
## installment FALSE FALSE  
## home\_ownership FALSE FALSE  
## annual\_inc FALSE FALSE  
## dti FALSE FALSE  
## delinq\_2yrs FALSE FALSE  
## fico\_range\_high FALSE FALSE  
## inq\_last\_6mths FALSE FALSE  
## open\_acc FALSE FALSE  
## pub\_rec FALSE FALSE  
## revol\_bal FALSE FALSE  
## total\_acc FALSE FALSE  
## out\_prncp FALSE FALSE  
## out\_prncp\_inv FALSE FALSE  
## total\_pymnt FALSE FALSE  
## total\_rec\_int FALSE FALSE  
## total\_rec\_late\_fee FALSE FALSE  
## recoveries FALSE FALSE  
## collection\_recovery\_fee FALSE FALSE  
## last\_fico\_range\_high FALSE FALSE  
## acc\_now\_delinq FALSE FALSE  
## inq\_last\_12m FALSE FALSE  
## num\_tl\_30dpd FALSE FALSE  
## num\_tl\_90g\_dpd\_24m FALSE FALSE  
## tot\_hi\_cred\_lim FALSE FALSE  
## num\_rev\_accts FALSE FALSE  
## is\_acct\_delinquent1 FALSE FALSE  
## 1 subsets of each size up to 28  
## Selection Algorithm: exhaustive  
## loan\_amnt term installment home\_ownership annual\_inc dti delinq\_2yrs  
## 1 ( 1 ) " " " " " " " " " " " " " "   
## 2 ( 1 ) " " " " " " " " " " " " " "   
## 3 ( 1 ) "\*" "\*" "\*" " " " " " " " "   
## 4 ( 1 ) "\*" "\*" "\*" " " " " " " " "   
## 5 ( 1 ) "\*" "\*" "\*" " " " " " " " "   
## 6 ( 1 ) "\*" "\*" "\*" " " " " " " " "   
## 7 ( 1 ) "\*" "\*" "\*" " " " " "\*" " "   
## 8 ( 1 ) "\*" "\*" "\*" " " " " "\*" " "   
## 9 ( 1 ) "\*" "\*" "\*" " " " " "\*" " "   
## 10 ( 1 ) "\*" "\*" "\*" " " " " "\*" " "   
## 11 ( 1 ) "\*" "\*" "\*" "\*" " " "\*" " "   
## 12 ( 1 ) "\*" "\*" "\*" "\*" " " "\*" "\*"   
## 13 ( 1 ) "\*" "\*" "\*" "\*" " " "\*" "\*"   
## 14 ( 1 ) "\*" "\*" "\*" "\*" " " "\*" "\*"   
## 15 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*"   
## 16 ( 1 ) "\*" "\*" "\*" "\*" " " "\*" "\*"   
## 17 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*"   
## 18 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*"   
## 19 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*"   
## 20 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*"   
## 21 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*"   
## 22 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*"   
## 23 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*"   
## 24 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*"   
## 25 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*"   
## 26 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*"   
## 27 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*"   
## 28 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*"   
## fico\_range\_high inq\_last\_6mths open\_acc pub\_rec revol\_bal total\_acc  
## 1 ( 1 ) "\*" " " " " " " " " " "   
## 2 ( 1 ) "\*" " " " " " " " " " "   
## 3 ( 1 ) " " " " " " " " " " " "   
## 4 ( 1 ) "\*" " " " " " " " " " "   
## 5 ( 1 ) "\*" " " " " " " " " " "   
## 6 ( 1 ) "\*" " " " " " " " " " "   
## 7 ( 1 ) "\*" " " " " " " " " " "   
## 8 ( 1 ) "\*" "\*" " " " " " " " "   
## 9 ( 1 ) "\*" "\*" " " " " " " " "   
## 10 ( 1 ) "\*" "\*" " " "\*" " " " "   
## 11 ( 1 ) "\*" "\*" " " "\*" " " " "   
## 12 ( 1 ) "\*" "\*" " " "\*" " " " "   
## 13 ( 1 ) "\*" "\*" " " "\*" " " " "   
## 14 ( 1 ) "\*" "\*" " " "\*" " " " "   
## 15 ( 1 ) "\*" "\*" " " "\*" " " " "   
## 16 ( 1 ) "\*" "\*" " " "\*" " " " "   
## 17 ( 1 ) "\*" "\*" " " "\*" " " " "   
## 18 ( 1 ) "\*" "\*" " " "\*" " " " "   
## 19 ( 1 ) "\*" "\*" " " "\*" " " " "   
## 20 ( 1 ) "\*" "\*" " " "\*" " " " "   
## 21 ( 1 ) "\*" "\*" " " "\*" " " " "   
## 22 ( 1 ) "\*" "\*" " " "\*" " " " "   
## 23 ( 1 ) "\*" "\*" " " "\*" " " "\*"   
## 24 ( 1 ) "\*" "\*" " " "\*" " " "\*"   
## 25 ( 1 ) "\*" "\*" " " "\*" " " "\*"   
## 26 ( 1 ) "\*" "\*" "\*" "\*" " " "\*"   
## 27 ( 1 ) "\*" "\*" "\*" "\*" " " "\*"   
## 28 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*"   
## out\_prncp out\_prncp\_inv total\_pymnt total\_rec\_int total\_rec\_late\_fee  
## 1 ( 1 ) " " " " " " " " " "   
## 2 ( 1 ) " " " " " " "\*" " "   
## 3 ( 1 ) " " " " " " " " " "   
## 4 ( 1 ) " " " " " " " " " "   
## 5 ( 1 ) " " " " " " "\*" " "   
## 6 ( 1 ) " " " " " " "\*" " "   
## 7 ( 1 ) " " " " " " "\*" " "   
## 8 ( 1 ) " " " " " " "\*" " "   
## 9 ( 1 ) " " " " " " "\*" " "   
## 10 ( 1 ) " " " " " " "\*" " "   
## 11 ( 1 ) " " " " " " "\*" " "   
## 12 ( 1 ) " " " " " " "\*" " "   
## 13 ( 1 ) " " " " " " "\*" " "   
## 14 ( 1 ) " " " " " " "\*" " "   
## 15 ( 1 ) " " " " " " "\*" " "   
## 16 ( 1 ) "\*" "\*" " " "\*" " "   
## 17 ( 1 ) "\*" "\*" " " "\*" " "   
## 18 ( 1 ) "\*" "\*" " " "\*" " "   
## 19 ( 1 ) "\*" "\*" " " "\*" " "   
## 20 ( 1 ) "\*" "\*" " " "\*" " "   
## 21 ( 1 ) "\*" "\*" " " "\*" "\*"   
## 22 ( 1 ) "\*" "\*" "\*" "\*" "\*"   
## 23 ( 1 ) "\*" "\*" "\*" "\*" "\*"   
## 24 ( 1 ) "\*" "\*" "\*" "\*" "\*"   
## 25 ( 1 ) "\*" "\*" "\*" "\*" "\*"   
## 26 ( 1 ) "\*" "\*" "\*" "\*" "\*"   
## 27 ( 1 ) "\*" "\*" "\*" "\*" "\*"   
## 28 ( 1 ) "\*" "\*" "\*" "\*" "\*"   
## recoveries collection\_recovery\_fee last\_fico\_range\_high  
## 1 ( 1 ) " " " " " "   
## 2 ( 1 ) " " " " " "   
## 3 ( 1 ) " " " " " "   
## 4 ( 1 ) " " " " " "   
## 5 ( 1 ) " " " " " "   
## 6 ( 1 ) " " " " "\*"   
## 7 ( 1 ) " " " " "\*"   
## 8 ( 1 ) " " " " "\*"   
## 9 ( 1 ) " " " " "\*"   
## 10 ( 1 ) " " " " "\*"   
## 11 ( 1 ) " " " " "\*"   
## 12 ( 1 ) " " " " "\*"   
## 13 ( 1 ) " " " " "\*"   
## 14 ( 1 ) " " " " "\*"   
## 15 ( 1 ) " " " " "\*"   
## 16 ( 1 ) " " " " "\*"   
## 17 ( 1 ) " " " " "\*"   
## 18 ( 1 ) "\*" " " "\*"   
## 19 ( 1 ) "\*" " " "\*"   
## 20 ( 1 ) "\*" " " "\*"   
## 21 ( 1 ) "\*" " " "\*"   
## 22 ( 1 ) "\*" " " "\*"   
## 23 ( 1 ) "\*" " " "\*"   
## 24 ( 1 ) "\*" " " "\*"   
## 25 ( 1 ) "\*" "\*" "\*"   
## 26 ( 1 ) "\*" "\*" "\*"   
## 27 ( 1 ) "\*" "\*" "\*"   
## 28 ( 1 ) "\*" "\*" "\*"   
## acc\_now\_delinq inq\_last\_12m num\_tl\_30dpd num\_tl\_90g\_dpd\_24m  
## 1 ( 1 ) " " " " " " " "   
## 2 ( 1 ) " " " " " " " "   
## 3 ( 1 ) " " " " " " " "   
## 4 ( 1 ) " " " " " " " "   
## 5 ( 1 ) " " " " " " " "   
## 6 ( 1 ) " " " " " " " "   
## 7 ( 1 ) " " " " " " " "   
## 8 ( 1 ) " " " " " " " "   
## 9 ( 1 ) " " " " " " " "   
## 10 ( 1 ) " " " " " " " "   
## 11 ( 1 ) " " " " " " " "   
## 12 ( 1 ) " " " " " " " "   
## 13 ( 1 ) " " "\*" " " " "   
## 14 ( 1 ) " " "\*" " " " "   
## 15 ( 1 ) " " "\*" " " " "   
## 16 ( 1 ) " " "\*" " " " "   
## 17 ( 1 ) " " "\*" " " " "   
## 18 ( 1 ) " " "\*" " " " "   
## 19 ( 1 ) " " "\*" " " "\*"   
## 20 ( 1 ) " " "\*" " " "\*"   
## 21 ( 1 ) " " "\*" " " "\*"   
## 22 ( 1 ) " " "\*" " " "\*"   
## 23 ( 1 ) " " "\*" " " "\*"   
## 24 ( 1 ) "\*" "\*" " " "\*"   
## 25 ( 1 ) "\*" "\*" " " "\*"   
## 26 ( 1 ) "\*" "\*" " " "\*"   
## 27 ( 1 ) "\*" "\*" "\*" "\*"   
## 28 ( 1 ) "\*" "\*" "\*" "\*"   
## tot\_hi\_cred\_lim num\_rev\_accts is\_acct\_delinquent1  
## 1 ( 1 ) " " " " " "   
## 2 ( 1 ) " " " " " "   
## 3 ( 1 ) " " " " " "   
## 4 ( 1 ) " " " " " "   
## 5 ( 1 ) " " " " " "   
## 6 ( 1 ) " " " " " "   
## 7 ( 1 ) " " " " " "   
## 8 ( 1 ) " " " " " "   
## 9 ( 1 ) " " "\*" " "   
## 10 ( 1 ) " " "\*" " "   
## 11 ( 1 ) " " "\*" " "   
## 12 ( 1 ) " " "\*" " "   
## 13 ( 1 ) " " "\*" " "   
## 14 ( 1 ) " " "\*" "\*"   
## 15 ( 1 ) " " "\*" "\*"   
## 16 ( 1 ) " " "\*" "\*"   
## 17 ( 1 ) " " "\*" "\*"   
## 18 ( 1 ) " " "\*" "\*"   
## 19 ( 1 ) " " "\*" "\*"   
## 20 ( 1 ) "\*" "\*" "\*"   
## 21 ( 1 ) "\*" "\*" "\*"   
## 22 ( 1 ) "\*" "\*" "\*"   
## 23 ( 1 ) "\*" "\*" "\*"   
## 24 ( 1 ) "\*" "\*" "\*"   
## 25 ( 1 ) "\*" "\*" "\*"   
## 26 ( 1 ) "\*" "\*" "\*"   
## 27 ( 1 ) "\*" "\*" "\*"   
## 28 ( 1 ) "\*" "\*" "\*"

regular\_subset\_full\_summary$rsq

## [1] 0.1972473 0.3264305 0.4706522 0.5469314 0.5920637 0.6071163 0.6097479  
## [8] 0.6118065 0.6145510 0.6161798 0.6169453 0.6173691 0.6177940 0.6181047  
## [15] 0.6183851 0.6186869 0.6189722 0.6190940 0.6192078 0.6192787 0.6193367  
## [22] 0.6193688 0.6193889 0.6194045 0.6194099 0.6194111 0.6194112 0.6194112

Drop variables based on best subset selection for interest rate

ds\_lc = subset(ds\_lc, select = -c(open\_acc, revol\_bal, total\_pymnt, total\_rec\_late\_fee, recoveries, collection\_recovery\_fee, acc\_now\_delinq, num\_tl\_30dpd, total\_rec\_int, out\_prncp, out\_prncp\_inv, is\_acct\_delinquent))

train\_dataset\_subset <- sample(nrow(ds\_lc) \* 0.75)  
train\_dataset <- ds\_lc[train\_dataset\_subset, ]  
test\_dataset <- ds\_lc[-train\_dataset\_subset, ]  
nrow(train\_dataset)

## [1] 370582

nrow(test\_dataset)

## [1] 123528

test\_interest\_mean <- mean(test\_dataset$int\_rate)  
test\_interest\_mse <- mean((test\_dataset$int\_rate - test\_interest\_mean)^2)  
test\_interest\_mean

## [1] 12.55028

test\_interest\_mse

## [1] 25.85992

## Fit a linear model using least squares on the training set, and identify the test error obtained.

lm\_fit <- lm(int\_rate ~ . , data = train\_dataset)  
lm\_predictions <- predict(lm\_fit, test\_dataset)  
lm\_mse <- mean((lm\_predictions - test\_dataset$int\_rate)^2)  
lm\_mse

## [1] 10.04733

summary(lm\_fit)

##   
## Call:  
## lm(formula = int\_rate ~ ., data = train\_dataset)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -20.2426 -2.2936 -0.4903 1.9166 30.3014   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.081e+01 1.356e-01 227.208 < 2e-16 \*\*\*  
## loan\_amnt -1.233e-03 2.926e-06 -421.453 < 2e-16 \*\*\*  
## term 1.084e+01 2.112e-02 513.405 < 2e-16 \*\*\*  
## installment 4.029e-02 9.496e-05 424.253 < 2e-16 \*\*\*  
## home\_ownership 1.513e-01 7.024e-03 21.544 < 2e-16 \*\*\*  
## annual\_inc -1.273e-06 6.829e-08 -18.635 < 2e-16 \*\*\*  
## dti 1.579e-02 2.747e-04 57.471 < 2e-16 \*\*\*  
## delinq\_2yrs 2.473e-01 1.023e-02 24.170 < 2e-16 \*\*\*  
## fico\_range\_high -2.801e-02 1.967e-04 -142.426 < 2e-16 \*\*\*  
## inq\_last\_6mths 3.460e-01 8.780e-03 39.404 < 2e-16 \*\*\*  
## pub\_rec -5.291e-01 1.471e-02 -35.970 < 2e-16 \*\*\*  
## total\_acc 2.214e-03 7.873e-04 2.812 0.00492 \*\*   
## last\_fico\_range\_high -1.634e-02 1.372e-04 -119.058 < 2e-16 \*\*\*  
## inq\_last\_12m 6.445e-02 2.787e-03 23.123 < 2e-16 \*\*\*  
## num\_tl\_90g\_dpd\_24m -2.069e-01 1.771e-02 -11.679 < 2e-16 \*\*\*  
## tot\_hi\_cred\_lim -5.877e-07 3.941e-08 -14.911 < 2e-16 \*\*\*  
## num\_rev\_accts -3.988e-02 1.128e-03 -35.360 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.392 on 370565 degrees of freedom  
## Multiple R-squared: 0.5691, Adjusted R-squared: 0.5691   
## F-statistic: 3.059e+04 on 16 and 370565 DF, p-value: < 2.2e-16

All the features are important predictors.

## Fit a ridge regression model and identify the test error obtained.

train\_ds\_matrix <- model.matrix(int\_rate ~ ., data = train\_dataset)  
test\_ds\_matrix <- model.matrix(int\_rate ~ ., data = test\_dataset)  
  
grid <- 10 ^ seq(4, -2, length = 100)  
  
ridge\_reg\_model <- cv.glmnet(train\_ds\_matrix, train\_dataset$int\_rate, alpha = 0, lambda = grid, thresh = 1e-12)  
ridge\_reg\_predictions <- predict(ridge\_reg\_model, test\_ds\_matrix, s = ridge\_reg\_model$lambda.min)  
ridge\_mse <- mean((test\_dataset$int\_rate - ridge\_reg\_predictions)^2)  
ridge\_mse

## [1] 10.05937

coef(ridge\_reg\_model)

## 18 x 1 sparse Matrix of class "dgCMatrix"  
## 1  
## (Intercept) 3.362102e+01  
## (Intercept) .   
## loan\_amnt -1.081631e-03  
## term 9.930775e+00  
## installment 3.538804e-02  
## home\_ownership 1.539241e-01  
## annual\_inc -1.374030e-06  
## dti 1.676189e-02  
## delinq\_2yrs 2.634489e-01  
## fico\_range\_high -2.975805e-02  
## inq\_last\_6mths 3.638271e-01  
## pub\_rec -5.341624e-01  
## total\_acc 2.086194e-03  
## last\_fico\_range\_high -1.707393e-02  
## inq\_last\_12m 6.902534e-02  
## num\_tl\_90g\_dpd\_24m -2.205964e-01  
## tot\_hi\_cred\_lim -7.294197e-07  
## num\_rev\_accts -4.114118e-02

Coefficients of all the variables are significant, all the features are important predictors.

## Fit a Lasso model and identify the test error obtained.

lasso\_model <- cv.glmnet(train\_ds\_matrix, train\_dataset$int\_rate, alpha = 1, lambda = grid, thresh = 1e-12)  
lasso\_predictions <- predict(lasso\_model, test\_ds\_matrix, s = lasso\_model$lambda.min)  
lasso\_mse <- mean((test\_dataset$int\_rate - lasso\_predictions)^2)  
lasso\_rmse <- sqrt(mean((test\_dataset$int\_rate - lasso\_predictions)^2))  
lasso\_mse

## [1] 10.04506

coef(lasso\_model)

## 18 x 1 sparse Matrix of class "dgCMatrix"  
## 1  
## (Intercept) 3.250388e+01  
## (Intercept) .   
## loan\_amnt -1.135045e-03  
## term 1.022579e+01  
## installment 3.708659e-02  
## home\_ownership 1.372898e-01  
## annual\_inc -1.178067e-06  
## dti 1.566262e-02  
## delinq\_2yrs 1.871028e-01  
## fico\_range\_high -2.894892e-02  
## inq\_last\_6mths 3.374613e-01  
## pub\_rec -4.823774e-01  
## total\_acc .   
## last\_fico\_range\_high -1.671242e-02  
## inq\_last\_12m 6.107966e-02  
## num\_tl\_90g\_dpd\_24m -8.760583e-02  
## tot\_hi\_cred\_lim -5.991500e-07  
## num\_rev\_accts -3.572901e-02

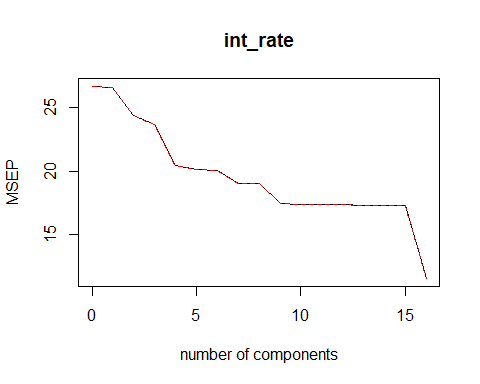
Apart for loan\_amnt and total\_acc all other variable/feature coefficients are statistically significant.

## PCR model

pcr\_model <- pcr(int\_rate ~ . , data = train\_dataset, scale=T, validation="CV")  
summary (pcr\_model)

## Data: X dimension: 370582 16   
## Y dimension: 370582 1  
## Fit method: svdpc  
## Number of components considered: 16  
##   
## VALIDATION: RMSEP  
## Cross-validated using 10 random segments.  
## (Intercept) 1 comps 2 comps 3 comps 4 comps 5 comps 6 comps  
## CV 5.168 5.156 4.943 4.868 4.519 4.483 4.48  
## adjCV 5.168 5.156 4.943 4.868 4.519 4.483 4.48  
## 7 comps 8 comps 9 comps 10 comps 11 comps 12 comps 13 comps  
## CV 4.363 4.361 4.179 4.169 4.168 4.163 4.16  
## adjCV 4.363 4.361 4.179 4.169 4.168 4.163 4.16  
## 14 comps 15 comps 16 comps  
## CV 4.158 4.157 3.392  
## adjCV 4.158 4.157 3.392  
##   
## TRAINING: % variance explained  
## 1 comps 2 comps 3 comps 4 comps 5 comps 6 comps 7 comps  
## X 18.3652 31.256 41.53 50.96 58.41 65.33 71.90  
## int\_rate 0.4747 8.531 11.30 23.53 24.76 24.87 28.73  
## 8 comps 9 comps 10 comps 11 comps 12 comps 13 comps 14 comps  
## X 77.79 83.36 88.09 91.36 94.08 96.57 98.62  
## int\_rate 28.81 34.62 34.92 34.95 35.11 35.22 35.26  
## 15 comps 16 comps  
## X 99.88 100.00  
## int\_rate 35.32 56.91

validationplot(pcr\_model, val.type = "MSEP")



1. The lowest cross-validation error occurs when M = 8 component are used.
2. 77.79% of the variation (or information) contained in the predictors are captured by 8 principal components.

Compute the test MSE.

pcr\_predictions <- predict(pcr\_model, test\_dataset, ncomp = 8)  
pcr\_mse <- mean((test\_dataset$int\_rate - pcr\_predictions)^2)  
pcr\_rmse <- sqrt(mean((test\_dataset$int\_rate - pcr\_predictions)^2))  
pcr\_mse

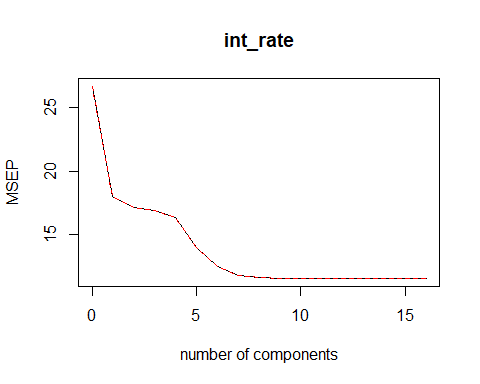
## [1] 17.00464

## PLS model

pls\_model <- plsr(int\_rate ~ . , data = train\_dataset, scale=T, validation="CV")  
summary (pls\_model)

## Data: X dimension: 370582 16   
## Y dimension: 370582 1  
## Fit method: kernelpls  
## Number of components considered: 16  
##   
## VALIDATION: RMSEP  
## Cross-validated using 10 random segments.  
## (Intercept) 1 comps 2 comps 3 comps 4 comps 5 comps 6 comps  
## CV 5.168 4.238 4.14 4.105 4.044 3.734 3.532  
## adjCV 5.168 4.238 4.14 4.105 4.044 3.734 3.532  
## 7 comps 8 comps 9 comps 10 comps 11 comps 12 comps 13 comps  
## CV 3.43 3.402 3.395 3.393 3.392 3.392 3.392  
## adjCV 3.43 3.402 3.395 3.393 3.392 3.392 3.392  
## 14 comps 15 comps 16 comps  
## CV 3.392 3.392 3.392  
## adjCV 3.392 3.392 3.392  
##   
## TRAINING: % variance explained  
## 1 comps 2 comps 3 comps 4 comps 5 comps 6 comps 7 comps  
## X 10.92 22.77 35.57 45.14 48.52 53.81 59.44  
## int\_rate 32.76 35.84 36.91 38.79 47.81 53.29 55.96  
## 8 comps 9 comps 10 comps 11 comps 12 comps 13 comps 14 comps  
## X 65.72 69.97 74.39 78.79 81.90 86.34 91.37  
## int\_rate 56.67 56.84 56.90 56.91 56.91 56.91 56.91  
## 15 comps 16 comps  
## X 95.80 100.00  
## int\_rate 56.91 56.91

validationplot(pls\_model, val.type = "MSEP")



1. The lowest cross-validation error occurs when M = 8 component are used, after M = 8 the error reduces marginally.
2. 65.72 % of the variation (or information) contained in the predictors are captured by 8 principal components.

Compute the test MSE.

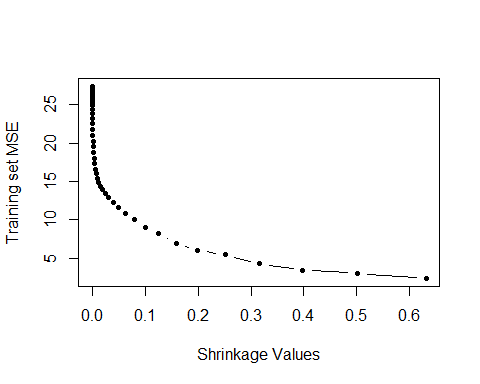
pls\_predictions <- predict(pls\_model, test\_dataset, ncomp = 8)  
pls\_mse <- mean((test\_dataset$int\_rate - pls\_predictions)^2)  
pls\_rmse <- sqrt(mean((test\_dataset$int\_rate - pls\_predictions)^2))  
pls\_mse

## [1] 10.1382

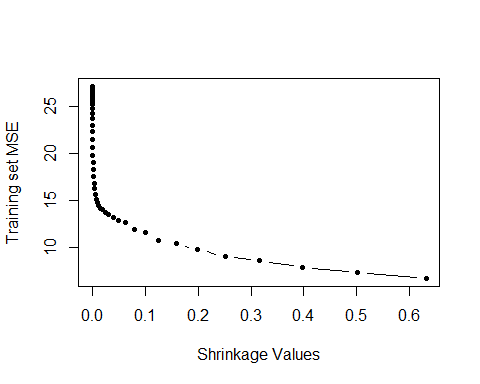
## Boosting

Changed the dataset size due to computing constraints

train\_dataset\_mini <- train\_dataset[1:1000,]  
test\_dataset\_mini <- test\_dataset[(1:1000),]  
  
pows <- seq(-10, -0.2, by = 0.1)  
lambdas <- 10^pows  
training\_errors <- rep(NA, length(lambdas))  
  
  
for (i in 1:length(lambdas)) {  
 boosting\_model <- gbm(int\_rate ~ . , data = train\_dataset\_mini, distribution = "gaussian",   
 n.trees = 1000, shrinkage = lambdas[i])  
   
 training\_predictions <- predict(boosting\_model, train\_dataset\_mini, n.trees = 1000)  
 training\_errors[i] <- mean((training\_predictions - train\_dataset\_mini$int\_rate)^2)  
}  
  
plot(lambdas, training\_errors, xlab = "Shrinkage Values", ylab = "Training set MSE", type = "b", pch = 20)



test\_errors <- rep(NA, length(lambdas))  
for (i in 1:length(lambdas)) {  
 boosting\_model <- gbm(int\_rate ~ . , data = train\_dataset\_mini, distribution = "gaussian",   
 n.trees = 1000, shrinkage = lambdas[i])  
   
 test\_predictions <- predict(boosting\_model, test\_dataset\_mini, n.trees = 1000)  
 test\_errors[i] <- mean((test\_predictions - test\_dataset\_mini$int\_rate)^2)  
}  
  
plot(lambdas, test\_errors, xlab = "Shrinkage Values", ylab = "Training set MSE", type = "b", pch = 20)



min(test\_errors)

## [1] 6.717623

boosting\_min\_test\_err <- min(test\_errors)  
boosting\_min\_test\_err\_at <- lambdas[which.min(test\_errors)]

Both regression approaches lm and lasso have higher MSE compared to that of boosting.

## Bagging

random\_forest\_model <- randomForest(int\_rate ~ . , data = train\_dataset\_mini, ntree = 500, mtry = ncol(train\_dataset\_mini)-1)  
random\_forest\_predictions <- predict(random\_forest\_model, test\_dataset\_mini)  
  
random\_forest\_test\_mse <- mean((random\_forest\_predictions - test\_dataset\_mini$int\_rate)^2)  
random\_forest\_test\_mse

## [1] 13.30536

Test MSE for bagging is 13.3053567 which is better than 6.7176233 which is best MSE from boosting

Compare R2 values for various models

lm\_test\_r2 <- (1 - (lm\_mse/test\_interest\_mse))  
ridge\_test\_r2 <- (1 - (ridge\_mse/test\_interest\_mse))  
lasso\_test\_r2 <- (1 - (lasso\_mse/test\_interest\_mse))  
pcr\_test\_r2 <- (1 - (pcr\_mse/test\_interest\_mse))  
pls\_test\_r2 <- (1 - (pls\_mse/test\_interest\_mse))  
boosting\_min\_test\_err\_r2 <- (1 - (boosting\_min\_test\_err/test\_interest\_mse))  
random\_forest\_r2 <- (1 - (random\_forest\_test\_mse/test\_interest\_mse))  
  
cat("R square with linear model : ", lm\_test\_r2, "\n")

## R square with linear model : 0.611471

cat("R square with ridge model : ", ridge\_test\_r2, "\n")

## R square with ridge model : 0.6110055

cat("R square with lasso model : ", lasso\_test\_r2, "\n")

## R square with lasso model : 0.6115587

cat("R square with pcr : ", pcr\_test\_r2, "\n")

## R square with pcr : 0.3424327

cat("R square with pls : ", pls\_test\_r2, "\n")

## R square with pls : 0.607957

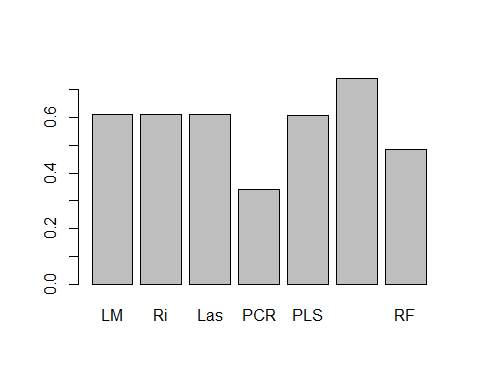
cat("R square with boosting : ", boosting\_min\_test\_err\_r2, "\n")

## R square with boosting : 0.7402303

cat("R square with random forest :", random\_forest\_r2, "\n")

## R square with random forest : 0.4854834

barplot(c(lm\_test\_r2, ridge\_test\_r2, lasso\_test\_r2, pcr\_test\_r2, pls\_test\_r2, boosting\_min\_test\_err\_r2, random\_forest\_r2),  
 names.arg = c("LM", "Ri", "Las", "PCR", "PLS", "Boost", "RF"))



All models have R Square values near to 0.6. All models except PCR predict interest rate, pcr has less accuracy than others.

# Unsupervised Learning

## PCA

### Data preparation for PCA

Some of the variables contain majority constant values so removing those variables.

ds\_lc\_pca <- ds\_lc\_after\_correlation  
ds\_lc\_pca = subset(ds\_lc\_pca, select = -c(is\_acct\_delinquent,pub\_rec, total\_rec\_late\_fee, recoveries, collection\_recovery\_fee, acc\_now\_delinq,  
 num\_tl\_30dpd,num\_tl\_90g\_dpd\_24m))  
ds\_lc\_pca\_subset <- sample(nrow(ds\_lc), 1000)  
ds\_lc\_pca = ds\_lc\_pca[ds\_lc\_pca\_subset, ]  
str(ds\_lc\_pca)

## Classes 'data.table' and 'data.frame': 1000 obs. of 21 variables:  
## $ loan\_amnt : int 26000 4200 5000 12500 40000 6000 20000 19200 7000 7000 ...  
## $ term : int 2 1 1 2 1 1 1 2 1 1 ...  
## $ int\_rate : num 8.81 15.02 5.31 11.8 6.46 ...  
## $ installment : num 537 146 151 277 1225 ...  
## $ home\_ownership : int 2 2 2 4 4 4 2 2 2 4 ...  
## $ annual\_inc : num 95000 37000 57000 34000 140000 26000 74000 30000 64000 60000 ...  
## $ dti : num 28.3 26.9 18.6 23.9 13.9 ...  
## $ delinq\_2yrs : int 0 0 0 0 0 0 1 0 0 0 ...  
## $ fico\_range\_high : int 719 694 684 674 789 704 719 679 799 734 ...  
## $ inq\_last\_6mths : int 0 2 0 1 0 0 0 0 0 0 ...  
## $ open\_acc : int 12 8 19 14 12 7 8 13 10 7 ...  
## $ revol\_bal : int 30377 8831 6162 11912 29519 3323 7274 21661 2394 5150 ...  
## $ total\_acc : int 27 14 44 15 17 10 20 16 15 9 ...  
## $ out\_prncp : num 24953 3524 3823 12034 35711 ...  
## $ out\_prncp\_inv : num 24953 3524 3823 12034 35711 ...  
## $ total\_pymnt : num 1599 1019 1361 810 4887 ...  
## $ total\_rec\_int : num 552 343 184 344 597 ...  
## $ last\_fico\_range\_high: int 694 709 689 754 834 749 684 594 774 734 ...  
## $ inq\_last\_12m : int 1 4 2 1 0 1 0 1 5 0 ...  
## $ tot\_hi\_cred\_lim : num 243770 100110 243201 58433 113330 ...  
## $ num\_rev\_accts : int 11 7 18 8 12 4 6 11 10 5 ...  
## - attr(\*, ".internal.selfref")=<externalptr>

pr.out <- prcomp (ds\_lc\_pca, scale =TRUE)  
names(pr.out)

## [1] "sdev" "rotation" "center" "scale" "x"

pr.out$center

## loan\_amnt term int\_rate   
## 16008.50000 1.30100 12.60280   
## installment home\_ownership annual\_inc   
## 465.10626 2.84800 83613.67016   
## dti delinq\_2yrs fico\_range\_high   
## 18.88726 0.24300 710.78600   
## inq\_last\_6mths open\_acc revol\_bal   
## 0.44600 11.49400 15243.13600   
## total\_acc out\_prncp out\_prncp\_inv   
## 23.11200 11959.99000 11958.80823   
## total\_pymnt total\_rec\_int last\_fico\_range\_high   
## 4924.32971 1146.53324 708.60500   
## inq\_last\_12m tot\_hi\_cred\_lim num\_rev\_accts   
## 1.87800 188398.34400 13.05200

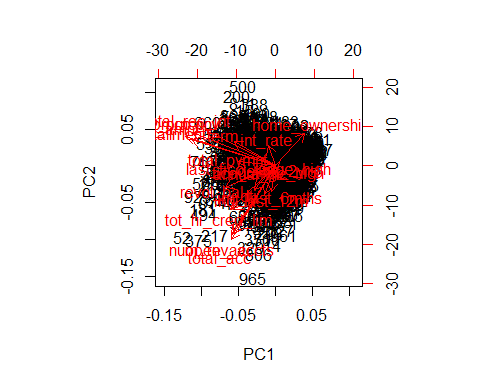
pr.out$scale

## loan\_amnt term int\_rate   
## 1.018067e+04 4.589222e-01 5.195439e+00   
## installment home\_ownership annual\_inc   
## 2.869426e+02 9.368958e-01 1.381947e+05   
## dti delinq\_2yrs fico\_range\_high   
## 1.453817e+01 7.632388e-01 3.625715e+01   
## inq\_last\_6mths open\_acc revol\_bal   
## 7.626701e-01 5.884947e+00 1.652522e+04   
## total\_acc out\_prncp out\_prncp\_inv   
## 1.207490e+01 9.470860e+03 9.471194e+03   
## total\_pymnt total\_rec\_int last\_fico\_range\_high   
## 5.574087e+03 1.113976e+03 5.187205e+01   
## inq\_last\_12m tot\_hi\_cred\_lim num\_rev\_accts   
## 2.380501e+00 1.815526e+05 7.584116e+00

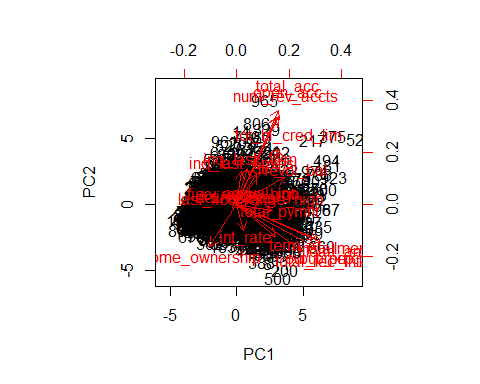
pr.out$rotation

## PC1 PC2 PC3 PC4  
## loan\_amnt -0.404699929 0.18020048 0.06678035 -0.080895123  
## term -0.186613709 0.15203522 -0.08107637 0.114633821  
## int\_rate -0.033565617 0.12526072 -0.50534983 0.037318868  
## installment -0.379406457 0.16345094 0.01617357 -0.130678092  
## home\_ownership 0.131512330 0.20155405 -0.15299716 0.262204125  
## annual\_inc -0.091650584 -0.04162478 0.03713122 -0.102728842  
## dti -0.095042601 -0.14307942 -0.13643688 0.291837940  
## delinq\_2yrs 0.006351724 -0.02870896 -0.15690583 -0.091723167  
## fico\_range\_high -0.024469835 -0.03803760 0.49695460 -0.069146145  
## inq\_last\_6mths -0.019821405 -0.15960026 -0.22681187 -0.435330531  
## open\_acc -0.197328916 -0.42357848 -0.11024346 0.210556838  
## revol\_bal -0.226950306 -0.13251028 -0.01008077 0.118839507  
## total\_acc -0.198061652 -0.45436394 -0.03850568 0.133698807  
## out\_prncp -0.375120180 0.20595773 0.03965322 0.104270758  
## out\_prncp\_inv -0.375116517 0.20600901 0.03963280 0.104225125  
## total\_pymnt -0.169147436 0.02610464 0.03691171 -0.324586338  
## total\_rec\_int -0.312021022 0.21910713 -0.19751123 -0.015892901  
## last\_fico\_range\_high -0.057360458 -0.01846376 0.48400545 0.007098865  
## inq\_last\_12m -0.050167978 -0.17293672 -0.21352554 -0.543631132  
## tot\_hi\_cred\_lim -0.202799506 -0.26658282 0.15985064 -0.214243540  
## num\_rev\_accts -0.191188873 -0.41611122 -0.06509433 0.222492658  
## PC5 PC6 PC7 PC8  
## loan\_amnt -0.06139676 -0.116352185 0.017284942 -0.06159791  
## term -0.15733747 0.325812397 -0.043924111 0.59088565  
## int\_rate -0.04101181 0.046759587 -0.151282133 0.28397441  
## installment -0.02196542 -0.236160134 0.008929558 -0.22260989  
## home\_ownership -0.34674840 -0.202580115 0.398241402 -0.15684077  
## annual\_inc 0.27707736 -0.011876690 0.609940237 0.17725672  
## dti -0.11140473 -0.001041425 -0.536912016 -0.13218413  
## delinq\_2yrs 0.33782795 -0.182468046 0.062543201 0.39762808  
## fico\_range\_high -0.27179530 0.109228887 -0.060384006 0.17791478  
## inq\_last\_6mths -0.31267261 0.249842731 0.092306279 -0.15836615  
## open\_acc -0.13368475 -0.033281021 0.141243798 0.01824134  
## revol\_bal 0.33683637 -0.100570556 -0.079790740 -0.23130558  
## total\_acc -0.10141120 -0.051537985 0.109208320 0.12755880  
## out\_prncp 0.04438885 0.262211861 0.095775998 -0.14973023  
## out\_prncp\_inv 0.04441292 0.262209083 0.095746061 -0.14970178  
## total\_pymnt -0.20409541 -0.660990401 -0.126660636 0.16292934  
## total\_rec\_int -0.09341898 -0.083874757 -0.073620040 0.16528625  
## last\_fico\_range\_high -0.26536114 0.035841589 -0.025566055 0.19122732  
## inq\_last\_12m -0.16444403 0.248231588 0.005543941 -0.12957679  
## tot\_hi\_cred\_lim 0.38293279 0.100106576 -0.187461819 0.07815704  
## num\_rev\_accts -0.18464455 -0.072578021 0.161113108 0.01618363  
## PC9 PC10 PC11 PC12  
## loan\_amnt 0.03521422 -0.003892304 0.01405830 -0.015113004  
## term -0.22990875 0.106628560 -0.23851152 -0.185190753  
## int\_rate -0.16086229 0.084704617 0.03602671 0.111332761  
## installment 0.09256730 -0.019171976 0.10497550 0.060336169  
## home\_ownership -0.02114736 -0.160557114 -0.33314468 -0.315371034  
## annual\_inc -0.45226691 -0.446610034 0.24818631 0.028607519  
## dti -0.12547269 -0.682865947 0.15040171 -0.109033532  
## delinq\_2yrs 0.67881056 -0.349643731 -0.22853207 -0.080475438  
## fico\_range\_high -0.04266187 -0.182435183 -0.09966461 0.119274833  
## inq\_last\_6mths 0.02205749 -0.233770444 -0.39770143 0.531854456  
## open\_acc 0.01627425 0.002333720 0.04277161 0.020554926  
## revol\_bal -0.28198980 0.051323194 -0.66919743 -0.161681256  
## total\_acc 0.09675142 0.095526680 0.10774076 0.031082987  
## out\_prncp 0.16644302 -0.023806384 0.03848110 -0.005182926  
## out\_prncp\_inv 0.16638376 -0.023754050 0.03856030 -0.005203844  
## total\_pymnt -0.17980887 0.046371411 -0.02268331 -0.033657132  
## total\_rec\_int -0.01174644 0.024577753 0.09479697 0.147012027  
## last\_fico\_range\_high 0.11985731 -0.108849956 -0.11506166 -0.164957560  
## inq\_last\_12m 0.04112530 0.002899807 0.13947044 -0.674396732  
## tot\_hi\_cred\_lim -0.13673654 0.066380529 -0.07996168 -0.025652767  
## num\_rev\_accts 0.12600125 0.217789957 0.04745047 0.026348259  
## PC13 PC14 PC15 PC16  
## loan\_amnt -1.364306e-01 -0.145654901 0.008047845 0.172463958  
## term -2.612913e-01 -0.382885451 0.050806959 0.001981116  
## int\_rate 3.852519e-01 0.398145663 -0.084887642 0.438142475  
## installment 1.294611e-02 0.038357340 -0.021326310 0.226645940  
## home\_ownership -1.843914e-01 0.306030203 0.353199121 0.098043262  
## annual\_inc 1.155856e-01 -0.044325838 -0.091979334 0.030256537  
## dti -5.283538e-02 -0.112003869 0.034364739 0.051182531  
## delinq\_2yrs -5.705992e-02 0.007436389 -0.096941671 0.013973784  
## fico\_range\_high -3.266610e-01 0.492712127 -0.431825120 0.103887207  
## inq\_last\_6mths 4.118158e-02 -0.153028008 0.119071655 0.005789349  
## open\_acc -1.460252e-05 0.044654366 0.059899350 -0.409935583  
## revol\_bal 1.893872e-01 0.031430921 -0.336537756 -0.085278170  
## total\_acc -1.204136e-02 0.065799001 0.157730999 0.132767417  
## out\_prncp 1.927090e-02 -0.016180166 0.023892301 0.045139320  
## out\_prncp\_inv 1.918937e-02 -0.016245432 0.023833434 0.045194583  
## total\_pymnt -5.015986e-02 -0.198504664 0.040786799 0.013172654  
## total\_rec\_int 5.257907e-02 0.356399141 -0.019690254 -0.646171216  
## last\_fico\_range\_high 7.239136e-01 -0.034386055 0.209382919 -0.020573595  
## inq\_last\_12m 1.782394e-02 0.093818668 -0.176826956 -0.061153894  
## tot\_hi\_cred\_lim -1.895241e-01 0.327527676 0.593848581 0.127187114  
## num\_rev\_accts -1.988101e-02 -0.079180666 -0.269395264 0.260132439  
## PC17 PC18 PC19 PC20  
## loan\_amnt -0.109129882 -0.05199375 0.257025012 -0.7869870649  
## term -0.025074556 -0.01232773 0.170031252 0.1944229309  
## int\_rate -0.216012080 -0.05275770 -0.098847245 -0.0833198340  
## installment -0.164582738 -0.06082385 0.543001583 0.5508273726  
## home\_ownership 0.050357294 0.06479260 -0.003211831 0.0046758199  
## annual\_inc 0.058400203 0.05458255 0.007886795 0.0015553232  
## dti 0.090491371 0.06779153 -0.002931481 0.0079521020  
## delinq\_2yrs -0.039478770 0.02134818 -0.002052639 0.0006926754  
## fico\_range\_high -0.088598688 -0.08349425 -0.062633767 0.0190118625  
## inq\_last\_6mths 0.053121015 0.03267973 0.002537367 0.0008263378  
## open\_acc -0.709683725 -0.07272852 -0.026009857 -0.0111664648  
## revol\_bal 0.081155330 -0.11382081 0.014578031 -0.0032356981  
## total\_acc 0.435235355 -0.65265265 0.019705135 0.0037515954  
## out\_prncp -0.002857981 -0.02446933 -0.399518708 0.0940014354  
## out\_prncp\_inv -0.002794971 -0.02459400 -0.399612261 0.0939645867  
## total\_pymnt -0.027918952 -0.00581349 -0.500221612 0.1091594971  
## total\_rec\_int 0.375412221 0.16897767 0.129686693 -0.0388949703  
## last\_fico\_range\_high 0.008127793 0.08731712 0.092891518 -0.0255594478  
## inq\_last\_12m 0.023619559 -0.01429937 0.006985351 -0.0080899650  
## tot\_hi\_cred\_lim -0.024136709 0.27489497 -0.009456318 0.0048206619  
## num\_rev\_accts 0.211529351 0.64265603 0.001198721 0.0079075176  
## PC21  
## loan\_amnt 2.831799e-05  
## term 2.931467e-05  
## int\_rate 1.272982e-05  
## installment -4.026728e-05  
## home\_ownership -5.247971e-05  
## annual\_inc 9.947406e-06  
## dti -9.218044e-06  
## delinq\_2yrs -3.200334e-05  
## fico\_range\_high 6.506333e-06  
## inq\_last\_6mths -3.992927e-05  
## open\_acc -6.818028e-05  
## revol\_bal -1.641083e-05  
## total\_acc 5.474167e-05  
## out\_prncp 7.071253e-01  
## out\_prncp\_inv -7.070882e-01  
## total\_pymnt 6.309544e-05  
## total\_rec\_int -3.114628e-05  
## last\_fico\_range\_high -8.377064e-05  
## inq\_last\_12m 2.605887e-05  
## tot\_hi\_cred\_lim -3.838808e-05  
## num\_rev\_accts -4.419769e-05

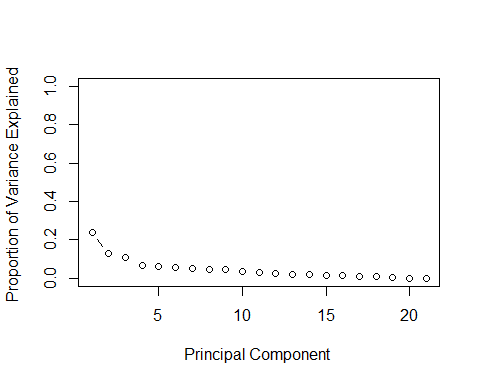
biplot(pr.out)



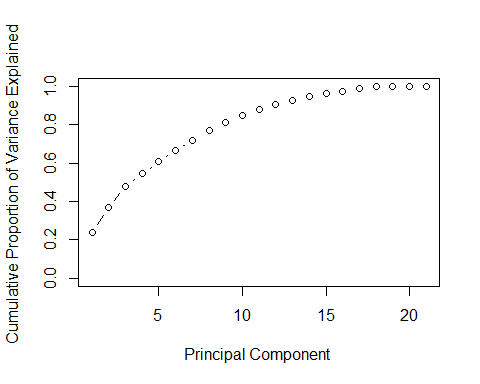
pr.out$rotation=-pr.out$rotation  
pr.out$x=-pr.out$x  
biplot (pr.out , scale =0)



pr.var =pr.out$sdev ^2  
pve=pr.var/sum(pr.var )  
plot(pve , xlab="Principal Component ", ylab="Proportion of Variance Explained ", ylim=c(0,1) ,type="b")

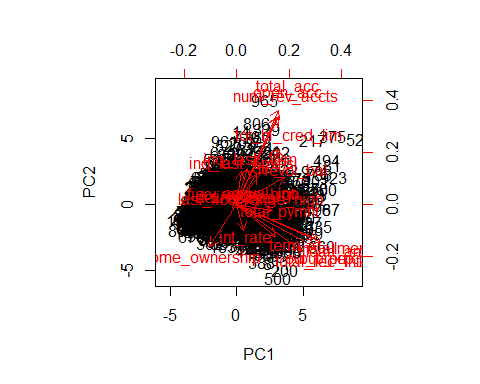


plot(cumsum (pve ), xlab="Principal Component ", ylab ="Cumulative Proportion of Variance Explained ", ylim=c(0,1) , type="b")

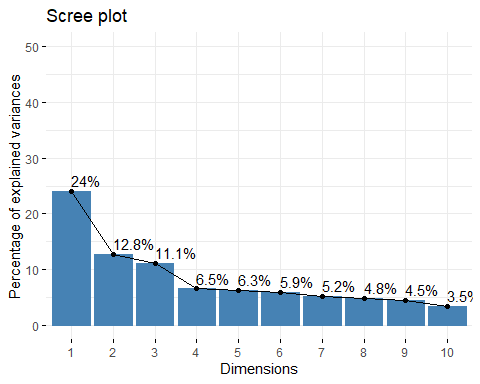


Component variance distribution

biplot (pr.out , scale =0)



fviz\_eig(pr.out, addlabels = TRUE, ylim = c(0, 50))



Based on the above plot first 6 components cover 66-70% of the variance, so baselining further analysis

Output for PCA

pr.out

## Standard deviations (1, .., p=21):  
## [1] 2.246588880 1.640319132 1.529343583 1.172776210 1.149066363 1.115109304  
## [7] 1.047077322 1.005254169 0.976611317 0.856100800 0.813507628 0.717400415  
## [13] 0.676036529 0.644723463 0.586561440 0.534034844 0.510760138 0.442475805  
## [19] 0.211060907 0.104309534 0.001055553  
##   
## Rotation (n x k) = (21 x 21):  
## PC1 PC2 PC3 PC4  
## loan\_amnt 0.404699929 -0.18020048 -0.06678035 0.080895123  
## term 0.186613709 -0.15203522 0.08107637 -0.114633821  
## int\_rate 0.033565617 -0.12526072 0.50534983 -0.037318868  
## installment 0.379406457 -0.16345094 -0.01617357 0.130678092  
## home\_ownership -0.131512330 -0.20155405 0.15299716 -0.262204125  
## annual\_inc 0.091650584 0.04162478 -0.03713122 0.102728842  
## dti 0.095042601 0.14307942 0.13643688 -0.291837940  
## delinq\_2yrs -0.006351724 0.02870896 0.15690583 0.091723167  
## fico\_range\_high 0.024469835 0.03803760 -0.49695460 0.069146145  
## inq\_last\_6mths 0.019821405 0.15960026 0.22681187 0.435330531  
## open\_acc 0.197328916 0.42357848 0.11024346 -0.210556838  
## revol\_bal 0.226950306 0.13251028 0.01008077 -0.118839507  
## total\_acc 0.198061652 0.45436394 0.03850568 -0.133698807  
## out\_prncp 0.375120180 -0.20595773 -0.03965322 -0.104270758  
## out\_prncp\_inv 0.375116517 -0.20600901 -0.03963280 -0.104225125  
## total\_pymnt 0.169147436 -0.02610464 -0.03691171 0.324586338  
## total\_rec\_int 0.312021022 -0.21910713 0.19751123 0.015892901  
## last\_fico\_range\_high 0.057360458 0.01846376 -0.48400545 -0.007098865  
## inq\_last\_12m 0.050167978 0.17293672 0.21352554 0.543631132  
## tot\_hi\_cred\_lim 0.202799506 0.26658282 -0.15985064 0.214243540  
## num\_rev\_accts 0.191188873 0.41611122 0.06509433 -0.222492658  
## PC5 PC6 PC7 PC8  
## loan\_amnt 0.06139676 0.116352185 -0.017284942 0.06159791  
## term 0.15733747 -0.325812397 0.043924111 -0.59088565  
## int\_rate 0.04101181 -0.046759587 0.151282133 -0.28397441  
## installment 0.02196542 0.236160134 -0.008929558 0.22260989  
## home\_ownership 0.34674840 0.202580115 -0.398241402 0.15684077  
## annual\_inc -0.27707736 0.011876690 -0.609940237 -0.17725672  
## dti 0.11140473 0.001041425 0.536912016 0.13218413  
## delinq\_2yrs -0.33782795 0.182468046 -0.062543201 -0.39762808  
## fico\_range\_high 0.27179530 -0.109228887 0.060384006 -0.17791478  
## inq\_last\_6mths 0.31267261 -0.249842731 -0.092306279 0.15836615  
## open\_acc 0.13368475 0.033281021 -0.141243798 -0.01824134  
## revol\_bal -0.33683637 0.100570556 0.079790740 0.23130558  
## total\_acc 0.10141120 0.051537985 -0.109208320 -0.12755880  
## out\_prncp -0.04438885 -0.262211861 -0.095775998 0.14973023  
## out\_prncp\_inv -0.04441292 -0.262209083 -0.095746061 0.14970178  
## total\_pymnt 0.20409541 0.660990401 0.126660636 -0.16292934  
## total\_rec\_int 0.09341898 0.083874757 0.073620040 -0.16528625  
## last\_fico\_range\_high 0.26536114 -0.035841589 0.025566055 -0.19122732  
## inq\_last\_12m 0.16444403 -0.248231588 -0.005543941 0.12957679  
## tot\_hi\_cred\_lim -0.38293279 -0.100106576 0.187461819 -0.07815704  
## num\_rev\_accts 0.18464455 0.072578021 -0.161113108 -0.01618363  
## PC9 PC10 PC11 PC12  
## loan\_amnt -0.03521422 0.003892304 -0.01405830 0.015113004  
## term 0.22990875 -0.106628560 0.23851152 0.185190753  
## int\_rate 0.16086229 -0.084704617 -0.03602671 -0.111332761  
## installment -0.09256730 0.019171976 -0.10497550 -0.060336169  
## home\_ownership 0.02114736 0.160557114 0.33314468 0.315371034  
## annual\_inc 0.45226691 0.446610034 -0.24818631 -0.028607519  
## dti 0.12547269 0.682865947 -0.15040171 0.109033532  
## delinq\_2yrs -0.67881056 0.349643731 0.22853207 0.080475438  
## fico\_range\_high 0.04266187 0.182435183 0.09966461 -0.119274833  
## inq\_last\_6mths -0.02205749 0.233770444 0.39770143 -0.531854456  
## open\_acc -0.01627425 -0.002333720 -0.04277161 -0.020554926  
## revol\_bal 0.28198980 -0.051323194 0.66919743 0.161681256  
## total\_acc -0.09675142 -0.095526680 -0.10774076 -0.031082987  
## out\_prncp -0.16644302 0.023806384 -0.03848110 0.005182926  
## out\_prncp\_inv -0.16638376 0.023754050 -0.03856030 0.005203844  
## total\_pymnt 0.17980887 -0.046371411 0.02268331 0.033657132  
## total\_rec\_int 0.01174644 -0.024577753 -0.09479697 -0.147012027  
## last\_fico\_range\_high -0.11985731 0.108849956 0.11506166 0.164957560  
## inq\_last\_12m -0.04112530 -0.002899807 -0.13947044 0.674396732  
## tot\_hi\_cred\_lim 0.13673654 -0.066380529 0.07996168 0.025652767  
## num\_rev\_accts -0.12600125 -0.217789957 -0.04745047 -0.026348259  
## PC13 PC14 PC15 PC16  
## loan\_amnt 1.364306e-01 0.145654901 -0.008047845 -0.172463958  
## term 2.612913e-01 0.382885451 -0.050806959 -0.001981116  
## int\_rate -3.852519e-01 -0.398145663 0.084887642 -0.438142475  
## installment -1.294611e-02 -0.038357340 0.021326310 -0.226645940  
## home\_ownership 1.843914e-01 -0.306030203 -0.353199121 -0.098043262  
## annual\_inc -1.155856e-01 0.044325838 0.091979334 -0.030256537  
## dti 5.283538e-02 0.112003869 -0.034364739 -0.051182531  
## delinq\_2yrs 5.705992e-02 -0.007436389 0.096941671 -0.013973784  
## fico\_range\_high 3.266610e-01 -0.492712127 0.431825120 -0.103887207  
## inq\_last\_6mths -4.118158e-02 0.153028008 -0.119071655 -0.005789349  
## open\_acc 1.460252e-05 -0.044654366 -0.059899350 0.409935583  
## revol\_bal -1.893872e-01 -0.031430921 0.336537756 0.085278170  
## total\_acc 1.204136e-02 -0.065799001 -0.157730999 -0.132767417  
## out\_prncp -1.927090e-02 0.016180166 -0.023892301 -0.045139320  
## out\_prncp\_inv -1.918937e-02 0.016245432 -0.023833434 -0.045194583  
## total\_pymnt 5.015986e-02 0.198504664 -0.040786799 -0.013172654  
## total\_rec\_int -5.257907e-02 -0.356399141 0.019690254 0.646171216  
## last\_fico\_range\_high -7.239136e-01 0.034386055 -0.209382919 0.020573595  
## inq\_last\_12m -1.782394e-02 -0.093818668 0.176826956 0.061153894  
## tot\_hi\_cred\_lim 1.895241e-01 -0.327527676 -0.593848581 -0.127187114  
## num\_rev\_accts 1.988101e-02 0.079180666 0.269395264 -0.260132439  
## PC17 PC18 PC19 PC20  
## loan\_amnt 0.109129882 0.05199375 -0.257025012 0.7869870649  
## term 0.025074556 0.01232773 -0.170031252 -0.1944229309  
## int\_rate 0.216012080 0.05275770 0.098847245 0.0833198340  
## installment 0.164582738 0.06082385 -0.543001583 -0.5508273726  
## home\_ownership -0.050357294 -0.06479260 0.003211831 -0.0046758199  
## annual\_inc -0.058400203 -0.05458255 -0.007886795 -0.0015553232  
## dti -0.090491371 -0.06779153 0.002931481 -0.0079521020  
## delinq\_2yrs 0.039478770 -0.02134818 0.002052639 -0.0006926754  
## fico\_range\_high 0.088598688 0.08349425 0.062633767 -0.0190118625  
## inq\_last\_6mths -0.053121015 -0.03267973 -0.002537367 -0.0008263378  
## open\_acc 0.709683725 0.07272852 0.026009857 0.0111664648  
## revol\_bal -0.081155330 0.11382081 -0.014578031 0.0032356981  
## total\_acc -0.435235355 0.65265265 -0.019705135 -0.0037515954  
## out\_prncp 0.002857981 0.02446933 0.399518708 -0.0940014354  
## out\_prncp\_inv 0.002794971 0.02459400 0.399612261 -0.0939645867  
## total\_pymnt 0.027918952 0.00581349 0.500221612 -0.1091594971  
## total\_rec\_int -0.375412221 -0.16897767 -0.129686693 0.0388949703  
## last\_fico\_range\_high -0.008127793 -0.08731712 -0.092891518 0.0255594478  
## inq\_last\_12m -0.023619559 0.01429937 -0.006985351 0.0080899650  
## tot\_hi\_cred\_lim 0.024136709 -0.27489497 0.009456318 -0.0048206619  
## num\_rev\_accts -0.211529351 -0.64265603 -0.001198721 -0.0079075176  
## PC21  
## loan\_amnt -2.831799e-05  
## term -2.931467e-05  
## int\_rate -1.272982e-05  
## installment 4.026728e-05  
## home\_ownership 5.247971e-05  
## annual\_inc -9.947406e-06  
## dti 9.218044e-06  
## delinq\_2yrs 3.200334e-05  
## fico\_range\_high -6.506333e-06  
## inq\_last\_6mths 3.992927e-05  
## open\_acc 6.818028e-05  
## revol\_bal 1.641083e-05  
## total\_acc -5.474167e-05  
## out\_prncp -7.071253e-01  
## out\_prncp\_inv 7.070882e-01  
## total\_pymnt -6.309544e-05  
## total\_rec\_int 3.114628e-05  
## last\_fico\_range\_high 8.377064e-05  
## inq\_last\_12m -2.605887e-05  
## tot\_hi\_cred\_lim 3.838808e-05  
## num\_rev\_accts 4.419769e-05

Choose first 6 components

pr.out$rotation[1:21,1:6]

## PC1 PC2 PC3 PC4  
## loan\_amnt 0.404699929 -0.18020048 -0.06678035 0.080895123  
## term 0.186613709 -0.15203522 0.08107637 -0.114633821  
## int\_rate 0.033565617 -0.12526072 0.50534983 -0.037318868  
## installment 0.379406457 -0.16345094 -0.01617357 0.130678092  
## home\_ownership -0.131512330 -0.20155405 0.15299716 -0.262204125  
## annual\_inc 0.091650584 0.04162478 -0.03713122 0.102728842  
## dti 0.095042601 0.14307942 0.13643688 -0.291837940  
## delinq\_2yrs -0.006351724 0.02870896 0.15690583 0.091723167  
## fico\_range\_high 0.024469835 0.03803760 -0.49695460 0.069146145  
## inq\_last\_6mths 0.019821405 0.15960026 0.22681187 0.435330531  
## open\_acc 0.197328916 0.42357848 0.11024346 -0.210556838  
## revol\_bal 0.226950306 0.13251028 0.01008077 -0.118839507  
## total\_acc 0.198061652 0.45436394 0.03850568 -0.133698807  
## out\_prncp 0.375120180 -0.20595773 -0.03965322 -0.104270758  
## out\_prncp\_inv 0.375116517 -0.20600901 -0.03963280 -0.104225125  
## total\_pymnt 0.169147436 -0.02610464 -0.03691171 0.324586338  
## total\_rec\_int 0.312021022 -0.21910713 0.19751123 0.015892901  
## last\_fico\_range\_high 0.057360458 0.01846376 -0.48400545 -0.007098865  
## inq\_last\_12m 0.050167978 0.17293672 0.21352554 0.543631132  
## tot\_hi\_cred\_lim 0.202799506 0.26658282 -0.15985064 0.214243540  
## num\_rev\_accts 0.191188873 0.41611122 0.06509433 -0.222492658  
## PC5 PC6  
## loan\_amnt 0.06139676 0.116352185  
## term 0.15733747 -0.325812397  
## int\_rate 0.04101181 -0.046759587  
## installment 0.02196542 0.236160134  
## home\_ownership 0.34674840 0.202580115  
## annual\_inc -0.27707736 0.011876690  
## dti 0.11140473 0.001041425  
## delinq\_2yrs -0.33782795 0.182468046  
## fico\_range\_high 0.27179530 -0.109228887  
## inq\_last\_6mths 0.31267261 -0.249842731  
## open\_acc 0.13368475 0.033281021  
## revol\_bal -0.33683637 0.100570556  
## total\_acc 0.10141120 0.051537985  
## out\_prncp -0.04438885 -0.262211861  
## out\_prncp\_inv -0.04441292 -0.262209083  
## total\_pymnt 0.20409541 0.660990401  
## total\_rec\_int 0.09341898 0.083874757  
## last\_fico\_range\_high 0.26536114 -0.035841589  
## inq\_last\_12m 0.16444403 -0.248231588  
## tot\_hi\_cred\_lim -0.38293279 -0.100106576  
## num\_rev\_accts 0.18464455 0.072578021

write.csv(pr.out$rotation[1:21,1:6], file = "data/pca\_output.csv", row.names=FALSE)

Based on the weights of variables across 5 components, here are the composition of variables across five components, in each of these componenets variables listed below are correlated with each other

PC1 => total\_pymnt, num\_rev\_accts, term, open\_acc, total\_acc, annual\_inc,revol\_bal, tot\_hi\_cred\_lim, total\_rec\_int, out\_prncp, installment, loan\_amnt

PC2 => inq\_last\_6mths, tot\_hi\_cred\_lim,open\_acc, total\_acc, num\_rev\_accts, total\_rec\_int, int\_rate,out\_prncp, out\_prncp\_inv, loan\_amnt, installment

PC3 => total\_acc,total\_rec\_int,delinq\_2yrs,dti,inq\_last\_12m,int\_rate,fico\_range\_high,last\_fico\_range\_high

PC4 => delinq\_2yrs,dti,revol\_bal, inq\_last\_12m, inq\_last\_6mths, total\_pymnt, fico\_range\_high

PC5 => annual\_inc,tot\_hi\_cred\_lim, delinq\_2yrs,total\_acc, total\_rec\_int, num\_rev\_accts, total\_pymnt, fico\_range\_high, open\_acc, home\_ownership, last\_fico\_range\_high,dti

PC6 => total\_pymnt, installment, home\_ownership, delinq\_2yrs, loan\_amnt, revol\_bal, inq\_last\_12m, inq\_last\_6mths, out\_prncp\_inv, out\_prncp, term

## K-Mean

### Data preparation for K-Mean

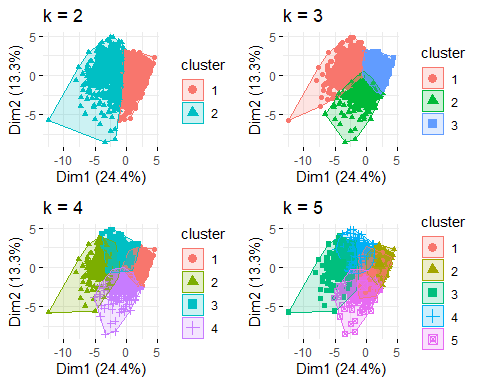
Some of the dataset variables contain constant values, so removing those variables.

ds\_lc\_kmean <- ds\_lc\_after\_correlation  
ds\_lc\_kmean = subset(ds\_lc\_kmean, select = -c(is\_acct\_delinquent,pub\_rec, total\_rec\_late\_fee, recoveries, collection\_recovery\_fee, acc\_now\_delinq,  
 num\_tl\_30dpd,num\_tl\_90g\_dpd\_24m))  
ds\_lc\_kmean\_subset <- sample(nrow(ds\_lc\_kmean), 2000)  
ds\_lc\_kmean = ds\_lc\_kmean[ds\_lc\_kmean\_subset, ]  
ds\_lc\_kmean\_scaled <- as.data.frame(scale(ds\_lc\_kmean))

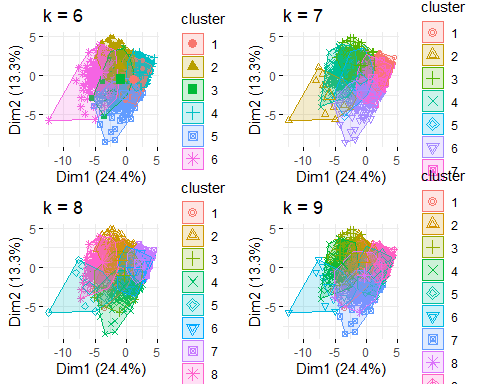
set.seed(123456)  
k2 <- kmeans(ds\_lc\_kmean\_scaled, centers = 2, nstart = 25)  
k3 <- kmeans(ds\_lc\_kmean\_scaled, centers = 3, nstart = 25)  
k4 <- kmeans(ds\_lc\_kmean\_scaled, centers = 4, nstart = 25)  
k5 <- kmeans(ds\_lc\_kmean\_scaled, centers = 5, nstart = 25)  
k6 <- kmeans(ds\_lc\_kmean\_scaled, centers = 6, nstart = 25)  
k7 <- kmeans(ds\_lc\_kmean\_scaled, centers = 7, nstart = 25)  
k8 <- kmeans(ds\_lc\_kmean\_scaled, centers = 8, nstart = 25)  
k9 <- kmeans(ds\_lc\_kmean\_scaled, centers = 9, nstart = 25)

## Warning: did not converge in 10 iterations

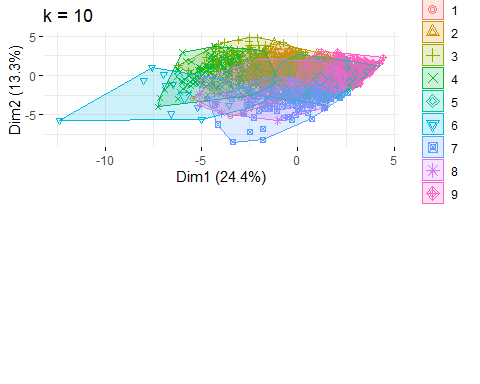
# plots to compare  
p2 <- fviz\_cluster(k2, geom = "point", data = ds\_lc\_kmean\_scaled, ggtheme = theme\_minimal()) + ggtitle("k = 2")  
p3 <- fviz\_cluster(k3, geom = "point", data = ds\_lc\_kmean\_scaled, ggtheme = theme\_minimal()) + ggtitle("k = 3")  
p4 <- fviz\_cluster(k4, geom = "point", data = ds\_lc\_kmean\_scaled, ggtheme = theme\_minimal()) + ggtitle("k = 4")  
p5 <- fviz\_cluster(k5, geom = "point", data = ds\_lc\_kmean\_scaled, ggtheme = theme\_minimal()) + ggtitle("k = 5")  
p6 <- fviz\_cluster(k6, geom = "point", data = ds\_lc\_kmean\_scaled, ggtheme = theme\_minimal()) + ggtitle("k = 6")  
p7 <- fviz\_cluster(k7, geom = "point", data = ds\_lc\_kmean\_scaled, ggtheme = theme\_minimal()) + ggtitle("k = 7")  
p8 <- fviz\_cluster(k8, geom = "point", data = ds\_lc\_kmean\_scaled, ggtheme = theme\_minimal()) + ggtitle("k = 8")  
p9 <- fviz\_cluster(k9, geom = "point", data = ds\_lc\_kmean\_scaled, ggtheme = theme\_minimal()) + ggtitle("k = 9")  
p10 <- fviz\_cluster(k9, geom = "point", data = ds\_lc\_kmean\_scaled, ggtheme = theme\_minimal()) + ggtitle("k = 10")  
  
  
grid.arrange(p2, p3, p4, p5, nrow = 2)



grid.arrange(p6, p7, p8, p9, nrow = 2)



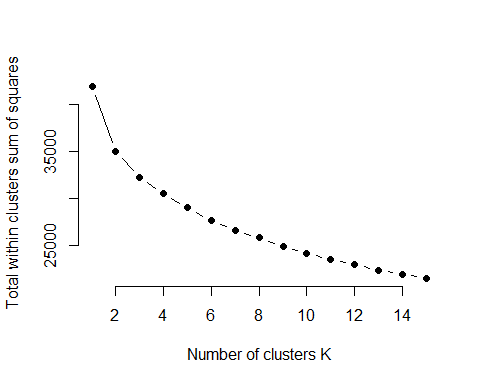
grid.arrange(p10, nrow = 2)



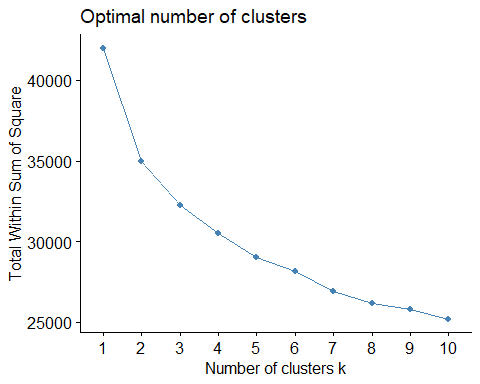
set.seed(123456)  
#Function to compute total within cluster sum of squares   
wss <- function(k) {  
 kmeans(ds\_lc\_kmean\_scaled, k, nstart = 25, iter.max = 10)$tot.withinss  
}  
  
#Compute and plot the within sum of squares (wss) for k = 1 to k = 10  
k.values <- 1:15  
  
#Extract wss for 2 - 10 clusters  
wss\_values <- map\_dbl(k.values, wss)

## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations

plot(k.values, wss\_values,  
 type = "b", pch = 19, frame = FALSE,  
 xlab = "Number of clusters K",  
 ylab = "Total within clusters sum of squares")



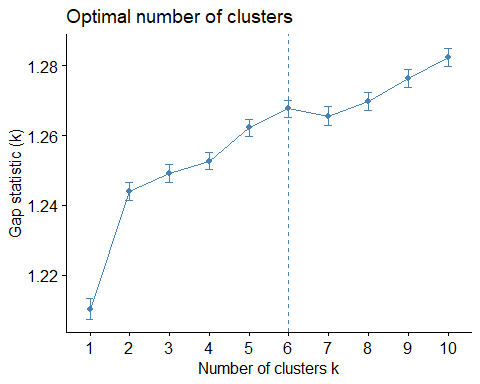
set.seed(123456)  
fviz\_nbclust(ds\_lc\_kmean\_scaled, kmeans, method = "wss")



set.seed(123456)  
gap\_stat <- clusGap(ds\_lc\_kmean\_scaled, FUN = kmeans, nstart = 25, K.max = 10, B = 50)

## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations

fviz\_gap\_stat(gap\_stat)

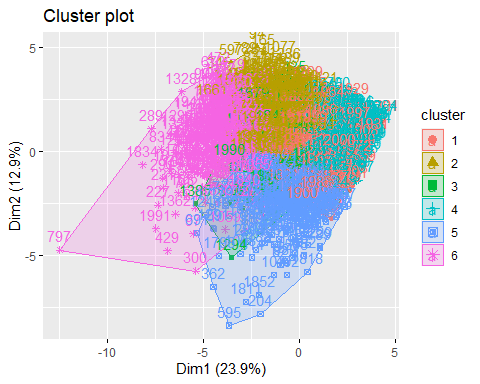


Based on the elbow method and GAP statistic method 6 number of clusters seem to be of optimal size.

ds\_lc\_kmean$cluster <- k6$cluster  
print(k6)

## K-means clustering with 6 clusters of sizes 385, 415, 79, 583, 276, 262  
##   
## Cluster means:  
## loan\_amnt term int\_rate installment home\_ownership annual\_inc  
## 1 -0.4725176 -0.4141416 -0.8859961 -0.4583495 -0.15258275 -0.14199933  
## 2 0.5178769 0.9945578 0.6635118 0.3760254 0.01306859 -0.10082667  
## 3 0.8141025 0.2955418 0.3801322 0.8740274 -0.28961529 0.29659526  
## 4 -0.7668994 -0.4783567 0.1736086 -0.6827118 0.45491700 -0.38195345  
## 5 -0.3658794 -0.1473310 0.1079979 -0.3159411 -0.37022685 0.05962438  
## 6 1.7205012 0.1637425 -0.3637426 1.6663609 -0.33142558 1.06604703  
## dti delinq\_2yrs fico\_range\_high inq\_last\_6mths open\_acc  
## 1 -0.107753324 -0.265258001 1.1099243 -0.28123477 -0.09160752  
## 2 0.099192773 0.003088562 -0.2163444 -0.01953781 -0.28273426  
## 3 0.005033529 -0.022284821 -0.1643261 0.21219888 0.26464480  
## 4 -0.192228828 -0.067991907 -0.5284090 -0.14342909 -0.58133751  
## 5 0.329587881 0.571874302 -0.3075367 0.71250097 1.28384594  
## 6 0.080250121 -0.059522702 0.2610168 -0.05118794 0.44379757  
## revol\_bal total\_acc out\_prncp out\_prncp\_inv total\_pymnt total\_rec\_int  
## 1 -0.23662954 -0.08919075 -0.4285220 -0.4283741 -0.22602013 -0.5872952  
## 2 -0.13576694 -0.22982396 0.6788951 0.6790522 -0.04453253 0.8227434  
## 3 0.04000611 0.27570628 -1.2710416 -1.2708873 3.75360987 0.2942777  
## 4 -0.35548563 -0.65152622 -0.6623895 -0.6625270 -0.42687576 -0.5178844  
## 5 0.20530467 1.30424876 -0.2554641 -0.2556548 -0.26755920 -0.2673487  
## 6 1.12545496 0.48779280 1.6806601 1.6806544 0.50258961 0.9051034  
## last\_fico\_range\_high inq\_last\_12m tot\_hi\_cred\_lim num\_rev\_accts  
## 1 0.8964925 -0.26930339 0.01255544 -0.1256809  
## 2 -0.1556551 -0.03265171 -0.14401821 -0.2717850  
## 3 0.1111011 0.36067217 0.39101949 0.1568482  
## 4 -0.5038835 -0.19328390 -0.53243981 -0.5205971  
## 5 -0.3489011 0.74611520 0.35215267 1.2802059  
## 6 0.3844696 -0.01719130 0.90557802 0.3777031  
##   
## Clustering vector:  
## [1] 5 4 4 1 1 4 2 2 4 5 6 4 2 2 4 4 4 2 2 3 2 4 4 4 6 4 4 1 5 2 2 5 2 1 2 2 2  
## [38] 1 2 4 4 5 6 4 4 5 1 3 5 2 4 6 4 5 4 5 6 6 5 3 4 6 1 4 2 2 4 5 4 1 6 1 1 2  
## [75] 4 2 2 1 1 2 4 5 1 1 3 6 2 1 2 1 2 4 4 2 4 6 6 2 1 2 5 5 2 1 2 2 5 4 1 1 2  
## [112] 4 4 4 4 4 4 1 6 4 4 4 4 1 1 5 1 6 6 6 4 6 4 6 4 2 6 6 4 1 4 6 2 1 2 4 1 1  
## [149] 6 6 2 6 4 4 4 4 4 2 4 4 2 5 5 1 2 6 5 3 4 2 2 1 2 2 2 2 1 5 2 1 3 2 4 5 2  
## [186] 4 3 1 5 2 4 4 6 6 2 4 4 6 5 2 5 1 5 5 2 2 4 4 6 1 2 6 4 6 4 1 5 5 2 5 1 4  
## [223] 2 6 1 4 6 6 5 1 6 1 4 1 1 2 1 5 5 5 5 5 4 2 5 4 1 6 2 2 2 2 4 5 2 6 1 4 4  
## [260] 1 5 4 6 1 2 2 5 1 5 1 4 1 4 3 1 4 6 2 1 5 2 4 2 2 5 4 2 2 6 5 6 3 2 5 6 5  
## [297] 2 4 1 6 5 1 1 5 4 6 2 4 4 1 1 2 6 2 6 4 5 3 3 6 4 6 5 1 6 2 4 1 4 4 4 5 6  
## [334] 4 2 2 4 6 2 6 1 5 1 4 2 4 2 4 6 4 6 4 4 4 2 5 4 4 6 4 6 5 6 1 3 2 2 4 5 4  
## [371] 4 2 4 1 3 2 2 4 4 4 4 5 2 4 4 6 4 1 2 5 1 4 1 1 1 4 6 6 4 1 1 5 4 4 5 4 5  
## [408] 2 1 1 6 1 1 4 2 4 1 2 1 2 4 3 2 4 4 6 2 4 6 4 4 4 3 2 3 4 1 4 3 4 5 3 2 4  
## [445] 4 1 4 4 4 2 5 4 4 5 4 6 4 1 1 4 2 5 5 6 2 4 6 5 6 1 1 6 1 6 4 2 3 5 1 4 2  
## [482] 2 5 1 4 2 4 4 2 5 1 1 4 2 5 2 1 1 1 4 6 4 6 4 1 5 1 2 4 4 4 2 1 5 4 4 4 5  
## [519] 5 2 4 4 6 1 4 1 4 6 4 4 2 4 6 6 4 5 2 2 2 2 2 1 5 4 4 5 6 2 1 3 4 5 2 4 4  
## [556] 6 5 2 6 4 1 3 4 4 1 4 5 1 2 1 6 1 2 1 2 1 2 6 4 1 1 4 4 4 4 4 1 5 1 4 4 2  
## [593] 4 1 5 6 2 2 4 4 6 2 1 1 1 4 4 4 2 6 5 1 2 4 2 6 1 3 6 5 4 1 2 4 2 2 1 2 4  
## [630] 1 2 1 2 1 2 2 4 6 6 4 6 3 4 2 1 4 3 2 2 6 1 1 5 4 1 4 6 2 1 4 5 1 4 6 6 5  
## [667] 1 6 4 4 1 6 5 2 3 2 3 2 2 6 2 1 3 1 2 1 4 3 2 4 5 4 4 5 5 1 5 5 6 4 2 2 5  
## [704] 6 1 6 4 4 1 2 2 1 4 3 6 4 1 4 2 4 4 6 6 1 3 1 2 1 2 5 4 4 1 6 1 6 4 4 4 1  
## [741] 4 4 6 4 4 4 6 1 5 2 6 2 5 1 4 4 2 1 1 2 5 4 2 5 4 5 3 6 6 4 4 3 5 5 4 2 1  
## [778] 5 5 1 1 4 4 3 6 6 4 2 2 4 1 6 4 2 5 2 6 4 5 1 1 3 5 2 4 2 1 2 4 5 2 4 6 5  
## [815] 4 2 4 2 6 1 4 4 6 1 6 4 5 1 4 2 2 4 1 6 2 2 2 4 6 4 1 6 1 5 6 4 1 4 6 1 4  
## [852] 6 1 2 1 4 4 6 6 1 2 4 2 2 4 2 5 2 5 4 5 5 4 6 1 1 4 2 1 4 2 5 4 4 4 4 6 4  
## [889] 4 6 4 2 1 4 5 4 2 6 5 4 4 1 4 5 4 1 5 1 2 4 4 1 1 6 1 2 4 4 4 4 6 4 5 3 6  
## [926] 4 4 1 1 1 1 5 5 1 1 4 2 4 1 4 3 1 1 4 1 5 4 2 5 2 6 2 4 4 4 1 2 3 1 4 2 2  
## [963] 5 1 5 5 2 1 4 2 4 2 2 4 1 1 5 5 4 2 4 5 4 1 1 4 4 5 4 5 2 2 2 6 2 1 2 5 4  
## [1000] 5 5 6 1 2 1 2 4 5 4 6 4 6 5 5 4 4 2 2 5 4 4 5 5 6 1 1 4 4 1 1 6 2 6 4 4 2  
## [1037] 4 3 4 6 4 4 2 1 1 2 1 2 1 2 4 2 2 4 4 4 4 3 6 4 2 4 4 2 2 6 4 2 1 2 2 4 1  
## [1074] 5 4 4 2 5 1 5 5 1 3 4 2 6 1 4 6 1 2 4 2 4 5 5 5 6 4 2 1 5 6 4 4 6 1 5 2 1  
## [1111] 6 4 2 2 4 4 4 6 4 1 2 2 3 2 2 4 1 5 4 4 2 5 2 2 6 4 4 4 2 3 1 2 6 6 4 1 4  
## [1148] 2 1 2 2 1 2 1 5 4 3 5 3 4 1 2 6 1 1 1 5 6 6 4 2 4 6 1 2 2 4 4 6 5 4 1 2 1  
## [1185] 2 4 2 5 2 1 2 4 4 4 4 1 1 1 1 2 1 6 4 3 3 2 4 5 1 3 6 4 2 6 5 6 6 4 5 4 2  
## [1222] 5 1 2 6 2 5 5 4 4 5 4 4 4 4 4 6 2 4 4 2 5 5 4 4 5 5 4 1 2 1 6 5 4 6 6 5 6  
## [1259] 4 1 1 4 5 2 4 4 5 2 1 6 5 2 2 4 4 1 5 5 4 5 4 5 4 5 1 1 4 1 6 6 5 1 1 3 1  
## [1296] 1 2 1 3 2 4 6 4 2 6 1 4 5 4 6 6 6 4 5 6 1 1 6 4 5 4 5 2 5 4 6 2 6 1 3 2 4  
## [1333] 5 4 4 5 1 2 4 1 4 6 1 2 4 6 4 5 4 2 4 6 4 4 6 4 2 1 5 3 5 6 1 4 1 2 4 1 1  
## [1370] 6 6 1 6 2 1 4 1 3 6 1 1 2 2 2 3 2 4 2 4 6 1 6 2 3 1 1 2 2 4 3 5 2 2 2 2 1  
## [1407] 6 3 4 4 4 2 6 2 4 5 2 3 4 1 2 5 1 5 4 5 4 4 5 4 5 4 5 4 5 4 5 6 1 4 2 6 2  
## [1444] 2 5 1 2 2 1 4 2 2 4 4 4 1 2 1 4 1 1 2 6 2 6 6 5 1 2 1 2 2 5 1 6 4 2 4 1 2  
## [1481] 4 2 1 6 2 3 2 4 6 4 4 5 1 4 2 5 2 4 1 2 4 6 3 1 5 4 6 4 1 5 1 4 1 1 4 1 2  
## [1518] 6 4 4 5 4 4 3 2 3 5 1 4 1 1 1 5 5 4 2 6 4 2 5 4 5 2 2 2 4 5 2 6 4 2 4 1 4  
## [1555] 1 6 5 1 5 5 2 3 4 4 2 4 4 5 5 2 1 1 1 4 6 4 6 2 1 4 3 2 4 3 4 5 1 4 1 5 2  
## [1592] 4 5 2 1 6 2 3 4 4 4 2 2 5 4 1 5 4 4 4 1 1 4 3 5 2 5 4 4 6 2 5 4 1 2 6 1 6  
## [1629] 2 4 1 1 2 5 2 4 6 3 6 4 6 4 4 1 2 5 5 2 1 6 4 1 1 1 4 4 2 4 2 6 2 1 4 4 5  
## [1666] 1 4 1 5 6 6 6 4 1 4 4 2 1 3 6 4 1 1 1 3 4 4 6 4 6 4 2 4 2 2 5 2 4 1 6 2 1  
## [1703] 3 1 5 4 6 2 2 1 3 4 4 5 4 6 1 6 6 1 5 4 2 4 5 2 4 5 4 1 4 4 1 1 2 2 2 4 6  
## [1740] 6 2 2 5 3 6 5 4 1 2 2 1 6 6 2 4 4 4 1 6 2 1 6 6 4 1 3 4 4 4 4 1 1 5 2 6 4  
## [1777] 5 2 6 4 2 4 1 4 4 5 5 2 4 5 2 5 4 5 4 4 2 5 2 4 5 6 4 2 2 2 1 4 2 4 5 6 4  
## [1814] 3 4 6 5 5 4 6 1 4 5 2 4 4 6 4 6 6 5 1 3 6 1 4 4 5 2 6 1 4 4 1 3 5 2 5 5 4  
## [1851] 4 5 2 4 4 4 1 1 4 4 1 1 5 2 4 6 2 6 1 2 6 6 2 4 2 1 4 4 4 4 4 1 5 6 1 2 5  
## [1888] 2 5 5 1 5 2 3 1 2 4 2 4 1 2 1 5 4 2 2 1 1 1 1 6 6 5 1 6 4 4 2 2 4 5 4 4 4  
## [1925] 2 2 1 4 1 1 1 4 1 2 1 2 4 6 5 4 2 4 2 2 4 1 4 1 6 6 2 1 2 5 6 4 3 5 4 2 1  
## [1962] 1 5 1 1 1 5 4 2 6 6 1 2 2 4 4 4 1 2 2 2 2 5 2 2 1 5 1 5 3 6 4 2 4 4 5 1 4  
## [1999] 1 1  
##   
## Within cluster sum of squares by cluster:  
## [1] 3714.713 6001.032 1401.022 4949.450 5146.968 6451.670  
## (between\_SS / total\_SS = 34.1 %)  
##   
## Available components:  
##   
## [1] "cluster" "centers" "totss" "withinss" "tot.withinss"  
## [6] "betweenss" "size" "iter" "ifault"

fviz\_cluster(k6, data = ds\_lc\_kmean)



Descriptive statistics at the cluster level

ds\_lc\_kmean %>%  
 group\_by(cluster) %>%  
 summarise\_all("mean")

## # A tibble: 6 x 22  
## cluster loan\_amnt term int\_rate installment home\_ownership annual\_inc dti  
## <int> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1 11246. 1.11 8.20 337. 2.74 70280. 17.9  
## 2 2 21158. 1.75 16.2 575. 2.89 72840. 20.8  
## 3 3 24123. 1.43 14.7 717. 2.61 97546. 19.5  
## 4 4 8300. 1.08 13.6 273. 3.30 55363. 16.7  
## 5 5 12314. 1.23 13.3 377. 2.53 82815. 24.1  
## 6 6 33194. 1.37 10.9 944. 2.57 145381. 20.6  
## # ... with 14 more variables: delinq\_2yrs <dbl>, fico\_range\_high <dbl>,  
## # inq\_last\_6mths <dbl>, open\_acc <dbl>, revol\_bal <dbl>, total\_acc <dbl>,  
## # out\_prncp <dbl>, out\_prncp\_inv <dbl>, total\_pymnt <dbl>,  
## # total\_rec\_int <dbl>, last\_fico\_range\_high <dbl>, inq\_last\_12m <dbl>,  
## # tot\_hi\_cred\_lim <dbl>, num\_rev\_accts <dbl>