HW2

Eva

6/23/2018

## reference   
##https://www.kaggle.com/farazrahman/lending-club-check-before-you-cheque/notebook  
## https://www.kaggle.com/erykwalczak/initial-loan-book-analysis  
##https://www.kaggle.com/dhanyajothimani/lending-loan-dataset-visualization  
  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(parsedate)  
library(lubridate)

##   
## Attaching package: 'lubridate'

## The following object is masked from 'package:base':  
##   
## date

library(DescTools)  
##question  
##Think about if/how you would process old features and what new features to be generated.  
##Build the best linear regression model to explain interest rate.  
rm(list=ls())  
setwd("~/evagit/BitTiger/Pro\_LendingClub")  
loan\_df= read.csv("loan.csv", stringsAsFactors = FALSE)  
glimpse(loan\_df)

## Observations: 887,379  
## Variables: 74  
## $ id <int> 1077501, 1077430, 1077175, 1076863...  
## $ member\_id <int> 1296599, 1314167, 1313524, 1277178...  
## $ loan\_amnt <dbl> 5000, 2500, 2400, 10000, 3000, 500...  
## $ funded\_amnt <dbl> 5000, 2500, 2400, 10000, 3000, 500...  
## $ funded\_amnt\_inv <dbl> 4975.00, 2500.00, 2400.00, 10000.0...  
## $ term <chr> " 36 months", " 60 months", " 36 m...  
## $ int\_rate <dbl> 10.65, 15.27, 15.96, 13.49, 12.69,...  
## $ installment <dbl> 162.87, 59.83, 84.33, 339.31, 67.7...  
## $ grade <chr> "B", "C", "C", "C", "B", "A", "C",...  
## $ sub\_grade <chr> "B2", "C4", "C5", "C1", "B5", "A4"...  
## $ emp\_title <chr> "", "Ryder", "", "AIR RESOURCES BO...  
## $ emp\_length <chr> "10+ years", "< 1 year", "10+ year...  
## $ home\_ownership <chr> "RENT", "RENT", "RENT", "RENT", "R...  
## $ annual\_inc <dbl> 24000.00, 30000.00, 12252.00, 4920...  
## $ verification\_status <chr> "Verified", "Source Verified", "No...  
## $ issue\_d <chr> "Dec-2011", "Dec-2011", "Dec-2011"...  
## $ loan\_status <chr> "Fully Paid", "Charged Off", "Full...  
## $ pymnt\_plan <chr> "n", "n", "n", "n", "n", "n", "n",...  
## $ url <chr> "https://www.lendingclub.com/brows...  
## $ desc <chr> " Borrower added on 12/22/11 > I ...  
## $ purpose <chr> "credit\_card", "car", "small\_busin...  
## $ title <chr> "Computer", "bike", "real estate b...  
## $ zip\_code <chr> "860xx", "309xx", "606xx", "917xx"...  
## $ addr\_state <chr> "AZ", "GA", "IL", "CA", "OR", "AZ"...  
## $ dti <dbl> 27.65, 1.00, 8.72, 20.00, 17.94, 1...  
## $ delinq\_2yrs <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...  
## $ earliest\_cr\_line <chr> "Jan-1985", "Apr-1999", "Nov-2001"...  
## $ inq\_last\_6mths <dbl> 1, 5, 2, 1, 0, 3, 1, 2, 2, 0, 2, 0...  
## $ mths\_since\_last\_delinq <dbl> NA, NA, NA, 35, 38, NA, NA, NA, NA...  
## $ mths\_since\_last\_record <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ open\_acc <dbl> 3, 3, 2, 10, 15, 9, 7, 4, 11, 2, 1...  
## $ pub\_rec <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...  
## $ revol\_bal <dbl> 13648, 1687, 2956, 5598, 27783, 79...  
## $ revol\_util <dbl> 83.70, 9.40, 98.50, 21.00, 53.90, ...  
## $ total\_acc <dbl> 9, 4, 10, 37, 38, 12, 11, 4, 13, 3...  
## $ initial\_list\_status <chr> "f", "f", "f", "f", "f", "f", "f",...  
## $ out\_prncp <dbl> 0.00, 0.00, 0.00, 0.00, 766.90, 0....  
## $ out\_prncp\_inv <dbl> 0.00, 0.00, 0.00, 0.00, 766.90, 0....  
## $ total\_pymnt <dbl> 5861.071, 1008.710, 3003.654, 1222...  
## $ total\_pymnt\_inv <dbl> 5831.78, 1008.71, 3003.65, 12226.3...  
## $ total\_rec\_prncp <dbl> 5000.00, 456.46, 2400.00, 10000.00...  
## $ total\_rec\_int <dbl> 861.07, 435.17, 603.65, 2209.33, 1...  
## $ total\_rec\_late\_fee <dbl> 0.00, 0.00, 0.00, 16.97, 0.00, 0.0...  
## $ recoveries <dbl> 0.00, 117.08, 0.00, 0.00, 0.00, 0....  
## $ collection\_recovery\_fee <dbl> 0.0000, 1.1100, 0.0000, 0.0000, 0....  
## $ last\_pymnt\_d <chr> "Jan-2015", "Apr-2013", "Jun-2014"...  
## $ last\_pymnt\_amnt <dbl> 171.62, 119.66, 649.91, 357.48, 67...  
## $ next\_pymnt\_d <chr> "", "", "", "", "Feb-2016", "", "F...  
## $ last\_credit\_pull\_d <chr> "Jan-2016", "Sep-2013", "Jan-2016"...  
## $ collections\_12\_mths\_ex\_med <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...  
## $ mths\_since\_last\_major\_derog <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ policy\_code <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...  
## $ application\_type <chr> "INDIVIDUAL", "INDIVIDUAL", "INDIV...  
## $ annual\_inc\_joint <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ dti\_joint <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ verification\_status\_joint <chr> "", "", "", "", "", "", "", "", ""...  
## $ acc\_now\_delinq <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...  
## $ tot\_coll\_amt <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ tot\_cur\_bal <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ open\_acc\_6m <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ open\_il\_6m <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ open\_il\_12m <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ open\_il\_24m <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ mths\_since\_rcnt\_il <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ total\_bal\_il <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ il\_util <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ open\_rv\_12m <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ open\_rv\_24m <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ max\_bal\_bc <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ all\_util <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ total\_rev\_hi\_lim <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ inq\_fi <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ total\_cu\_tl <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ inq\_last\_12m <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA...

sum(duplicated(loan\_df))

## [1] 0

## check if id and member\_id is unique => 887379   
length(unique(loan\_df$id))

## [1] 887379

length(unique(loan\_df$member\_id))

## [1] 887379

##select features with less than 20% missing value  
num.NA = sort(sapply(loan\_df, function(x){sum(is.na(x))}), decreasing=TRUE)  
remain.col = names(num.NA)[which(num.NA/nrow(loan\_df)<=0.8)]  
loan\_df=loan\_df[,remain.col]  
  
## check the columns with less than 4 unique values  
  
for(i in 1: length(loan\_df)){  
 if (dim(unique(loan\_df[i]))[1]<4){  
 print (names(loan\_df[i]))  
 print (table(loan\_df[i]))  
 }  
   
}

## [1] "term"  
##   
## 36 months 60 months   
## 621125 266254   
## [1] "verification\_status"  
##   
## Not Verified Source Verified Verified   
## 266750 329558 291071   
## [1] "pymnt\_plan"  
##   
## n y   
## 887369 10   
## [1] "initial\_list\_status"  
##   
## f w   
## 456848 430531   
## [1] "policy\_code"  
##   
## 1   
## 887379   
## [1] "application\_type"  
##   
## INDIVIDUAL JOINT   
## 886868 511

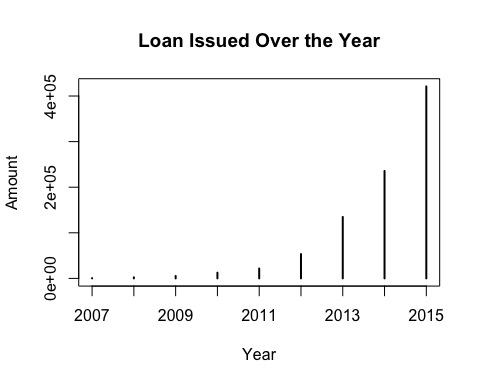
##following 3 features need to be concerned.  
##exclude "pymnt\_plan" and "policy\_code" from the dataset  
##we may keep the "application\_type" for further discussion  
  
# "pymnt\_plan"  
#   
# n y   
# 887369 10   
  
# [1] "policy\_code"  
#   
# 1   
# 887379   
  
# "application\_type"  
#   
# INDIVIDUAL JOINT   
# 886868 511   
  
  
##check all categorical variables  
loan\_df.type=sapply(loan\_df, class)  
loan\_df.categorical=loan\_df[,names(loan\_df.type[which(loan\_df.type=="character")])]  
str(loan\_df.categorical)

## 'data.frame': 887379 obs. of 23 variables:  
## $ term : chr " 36 months" " 60 months" " 36 months" " 36 months" ...  
## $ grade : chr "B" "C" "C" "C" ...  
## $ sub\_grade : chr "B2" "C4" "C5" "C1" ...  
## $ emp\_title : chr "" "Ryder" "" "AIR RESOURCES BOARD" ...  
## $ emp\_length : chr "10+ years" "< 1 year" "10+ years" "10+ years" ...  
## $ home\_ownership : chr "RENT" "RENT" "RENT" "RENT" ...  
## $ verification\_status : chr "Verified" "Source Verified" "Not Verified" "Source Verified" ...  
## $ issue\_d : chr "Dec-2011" "Dec-2011" "Dec-2011" "Dec-2011" ...  
## $ loan\_status : chr "Fully Paid" "Charged Off" "Fully Paid" "Fully Paid" ...  
## $ pymnt\_plan : chr "n" "n" "n" "n" ...  
## $ url : chr "https://www.lendingclub.com/browse/loanDetail.action?loan\_id=1077501" "https://www.lendingclub.com/browse/loanDetail.action?loan\_id=1077430" "https://www.lendingclub.com/browse/loanDetail.action?loan\_id=1077175" "https://www.lendingclub.com/browse/loanDetail.action?loan\_id=1076863" ...  
## $ desc : chr " Borrower added on 12/22/11 > I need to upgrade my business technologies.<br>" " Borrower added on 12/22/11 > I plan to use this money to finance the motorcycle i am looking at. I plan to ha"| \_\_truncated\_\_ "" " Borrower added on 12/21/11 > to pay for property tax (borrow from friend, need to pay back) & central A/C nee"| \_\_truncated\_\_ ...  
## $ purpose : chr "credit\_card" "car" "small\_business" "other" ...  
## $ title : chr "Computer" "bike" "real estate business" "personel" ...  
## $ zip\_code : chr "860xx" "309xx" "606xx" "917xx" ...  
## $ addr\_state : chr "AZ" "GA" "IL" "CA" ...  
## $ earliest\_cr\_line : chr "Jan-1985" "Apr-1999" "Nov-2001" "Feb-1996" ...  
## $ initial\_list\_status : chr "f" "f" "f" "f" ...  
## $ last\_pymnt\_d : chr "Jan-2015" "Apr-2013" "Jun-2014" "Jan-2015" ...  
## $ next\_pymnt\_d : chr "" "" "" "" ...  
## $ last\_credit\_pull\_d : chr "Jan-2016" "Sep-2013" "Jan-2016" "Jan-2015" ...  
## $ application\_type : chr "INDIVIDUAL" "INDIVIDUAL" "INDIVIDUAL" "INDIVIDUAL" ...  
## $ verification\_status\_joint: chr "" "" "" "" ...

##check unique values for each categorical variable  
sapply(loan\_df.categorical, function(x){ length(unique(x))})

## term grade   
## 2 7   
## sub\_grade emp\_title   
## 35 299273   
## emp\_length home\_ownership   
## 12 6   
## verification\_status issue\_d   
## 3 103   
## loan\_status pymnt\_plan   
## 10 2   
## url desc   
## 887379 124471   
## purpose title   
## 14 63146   
## zip\_code addr\_state   
## 935 51   
## earliest\_cr\_line initial\_list\_status   
## 698 2   
## last\_pymnt\_d next\_pymnt\_d   
## 99 101   
## last\_credit\_pull\_d application\_type   
## 104 2   
## verification\_status\_joint   
## 4

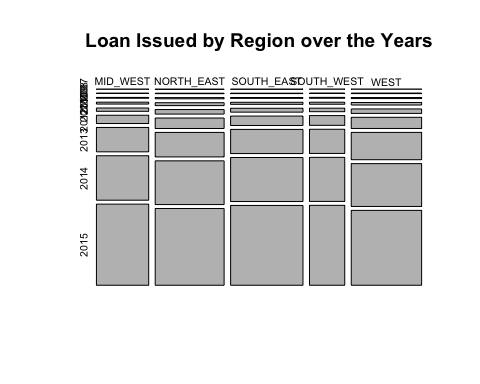
##delete the categorical values with too many unique values  
  
  
## after checking the features, deleting :   
#url,  
#desc => it's similar to purpose  
# loan\_df=loan\_df[,-which(colnames(loan\_df)%in%c("pymnt\_plan","policy\_code",'id', 'member\_id', 'url','desc', 'title','emp\_title','zip\_code'))]  
  
loan\_df <- within(loan\_df, rm("pymnt\_plan","policy\_code",'id', 'member\_id', 'url','desc', 'title','emp\_title','zip\_code'))  
  
##compute meaningful time features  
##convert date string to date, transform issue date by year  
  
loan\_df$issue\_year = sapply(loan\_df$issue\_d, function(x){substr(x, nchar(x)-3,nchar(x) )})  
plot(table(loan\_df$issue\_year), main="Loan Issued Over the Year", ylab = "Amount", xlab = "Year")



##too many levels  
##we can group the states into region by economic level / geolocation / average interest rate  
##group the states by region  
unique(loan\_df$addr\_state)

## [1] "AZ" "GA" "IL" "CA" "OR" "NC" "TX" "VA" "MO" "CT" "UT" "FL" "NY" "PA"  
## [15] "MN" "NJ" "KY" "OH" "SC" "RI" "LA" "MA" "WA" "WI" "AL" "CO" "KS" "NV"  
## [29] "AK" "MD" "WV" "VT" "MI" "DC" "SD" "NH" "AR" "NM" "MT" "HI" "WY" "OK"  
## [43] "DE" "MS" "TN" "IA" "NE" "ID" "IN" "ME" "ND"

west = c('CA', 'OR', 'UT','WA', 'CO', 'NV', 'AK', 'MT', 'HI', 'WY', 'ID')  
south\_west = c('AZ', 'TX', 'NM', 'OK')  
south\_east = c('GA', 'NC', 'VA', 'FL', 'KY', 'SC', 'LA', 'AL', 'WV', 'DC', 'AR', 'DE', 'MS', 'TN' )  
mid\_west = c('IL', 'MO', 'MN', 'OH', 'WI', 'KS', 'MI', 'SD', 'IA', 'NE', 'IN', 'ND')  
north\_east = c('CT', 'NY', 'PA', 'NJ', 'RI','MA', 'MD', 'VT', 'NH', 'ME')  
  
state\_to\_Reagion = function(s){  
 if (s %in% west){  
 return ("WEST")  
 }  
 if (s %in% south\_west){  
 return ("SOUTH\_WEST")  
 }  
 if (s %in% south\_east){  
 return ("SOUTH\_EAST")  
 }  
 if (s %in% mid\_west){  
 return ("MID\_WEST")  
 }  
 if (s %in% north\_east){  
 return ("NORTH\_EAST")  
 }  
 else{  
 return ("Missing\_Region")  
 }  
}  
  
loan\_df$Region = sapply(loan\_df$addr\_state, state\_to\_Reagion )  
  
plot(table(loan\_df$Region, loan\_df$issue\_year), main="Loan Issued by Region over the Years")



##remove addr\_state  
loan\_df=within(loan\_df, rm('addr\_state'))  
  
## check the employment length  
table(loan\_df$emp\_length)

##   
## < 1 year 1 year 10+ years 2 years 3 years 4 years 5 years   
## 70605 57095 291569 78870 70026 52529 55704   
## 6 years 7 years 8 years 9 years n/a   
## 42950 44594 43955 34657 44825

## convert employment length into numeric value  
emp\_length\_c\_to\_n=function(x){  
 if(x=="< 1 year"){  
 return(0.5)  
 }  
 if(x=="1 year"){  
 return(1)  
 }  
 if(x=="2 years"){  
 return(2)  
 }  
 if(x=="3 years"){  
 return(3)  
 }  
 if(x=="4 years"){  
 return(4)  
 }  
 if(x=="5 years"){  
 return(5)  
 }  
 if(x=="6 years"){  
 return(6)  
 }  
 if(x=="7 years"){  
 return(7)  
 }  
 if(x=="8 years"){  
 return(8)  
 }  
 if(x=="9 years"){  
 return(9)  
 }  
 if(x=="10+ years"){  
 return(10)  
 }  
 else{  
 return(0)  
 }  
   
}  
  
loan\_df$Employment\_Length = sapply(loan\_df$emp\_length, emp\_length\_c\_to\_n)  
summary(loan\_df$Employment\_Length)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 2.000 6.000 5.745 10.000 10.000

##delete some categorical variables  
loan\_df=loan\_df[,-which(colnames(loan\_df)%in%c("emp\_length"))]  
  
##delete the variables which will not be available at the time of deciding the intereste rate  
#remove the potential related responsive variable except "int\_rate"  
loan\_df=within(loan\_df, rm("funded\_amnt", "funded\_amnt\_inv", "out\_prncp", "out\_prncp\_inv", "total\_pymnt", "total\_pymnt\_inv", "total\_rec\_int", "total\_rec\_late\_fee", "total\_rec\_prncp", 'grade',"sub\_grade","loan\_status","last\_credit\_pull\_d","next\_pymnt\_d", "last\_pymnt\_amnt", "next\_pymnt\_d", "last\_pymnt\_d"))

## Warning in rm("funded\_amnt", "funded\_amnt\_inv", "out\_prncp",  
## "out\_prncp\_inv", : object 'next\_pymnt\_d' not found

##check the reminding categorical variables and convert them into factor  
loan\_df.type=sapply(loan\_df, class)  
loan\_df.categorical=loan\_df[,names(loan\_df.type[which(loan\_df.type=="character")])]  
str(loan\_df.categorical)

## 'data.frame': 887379 obs. of 11 variables:  
## $ term : chr " 36 months" " 60 months" " 36 months" " 36 months" ...  
## $ home\_ownership : chr "RENT" "RENT" "RENT" "RENT" ...  
## $ verification\_status : chr "Verified" "Source Verified" "Not Verified" "Source Verified" ...  
## $ issue\_d : chr "Dec-2011" "Dec-2011" "Dec-2011" "Dec-2011" ...  
## $ purpose : chr "credit\_card" "car" "small\_business" "other" ...  
## $ earliest\_cr\_line : chr "Jan-1985" "Apr-1999" "Nov-2001" "Feb-1996" ...  
## $ initial\_list\_status : chr "f" "f" "f" "f" ...  
## $ application\_type : chr "INDIVIDUAL" "INDIVIDUAL" "INDIVIDUAL" "INDIVIDUAL" ...  
## $ verification\_status\_joint: chr "" "" "" "" ...  
## $ issue\_year : chr "2011" "2011" "2011" "2011" ...  
## $ Region : chr "SOUTH\_WEST" "SOUTH\_EAST" "MID\_WEST" "WEST" ...

loan\_df$term=as.factor(loan\_df$term)  
loan\_df$home\_ownership=as.factor(loan\_df$home\_ownership)  
loan\_df$verification\_status=as.factor(loan\_df$verification\_status)  
loan\_df$purpose=as.factor(loan\_df$purpose)  
loan\_df$initial\_list\_status=as.factor(loan\_df$initial\_list\_status)  
loan\_df$application\_type=as.factor(loan\_df$application\_type)  
loan\_df$verification\_status\_joint=as.factor(loan\_df$verification\_status\_joint)  
loan\_df$Region=as.factor(loan\_df$Region)  
loan\_df$issue\_year=as.factor(loan\_df$issue\_year)  
loan\_df$issue\_d <- dmy(paste0("01-",loan\_df$issue\_d))  
loan\_df$earliest\_cr\_line <- dmy(paste0("01-",loan\_df$earliest\_cr\_line))

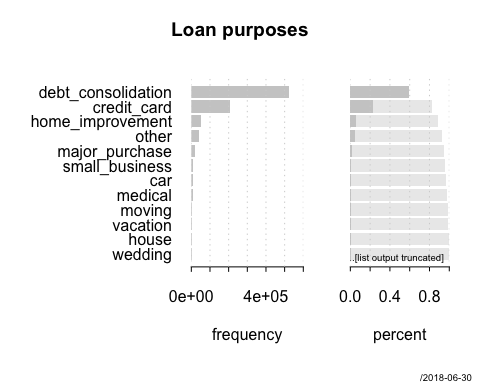
## Warning: 29 failed to parse.

loan\_df$earliest\_cr\_line\_year =as.factor(substring(loan\_df$earliest\_cr\_line, 1,4))  
glimpse(loan\_df)

## Observations: 887,379  
## Variables: 34  
## $ mths\_since\_last\_major\_derog <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ mths\_since\_last\_delinq <dbl> NA, NA, NA, 35, 38, NA, NA, NA, NA...  
## $ tot\_coll\_amt <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ tot\_cur\_bal <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ total\_rev\_hi\_lim <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ revol\_util <dbl> 83.70, 9.40, 98.50, 21.00, 53.90, ...  
## $ collections\_12\_mths\_ex\_med <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...  
## $ delinq\_2yrs <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...  
## $ inq\_last\_6mths <dbl> 1, 5, 2, 1, 0, 3, 1, 2, 2, 0, 2, 0...  
## $ open\_acc <dbl> 3, 3, 2, 10, 15, 9, 7, 4, 11, 2, 1...  
## $ pub\_rec <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...  
## $ total\_acc <dbl> 9, 4, 10, 37, 38, 12, 11, 4, 13, 3...  
## $ acc\_now\_delinq <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...  
## $ annual\_inc <dbl> 24000.00, 30000.00, 12252.00, 4920...  
## $ loan\_amnt <dbl> 5000, 2500, 2400, 10000, 3000, 500...  
## $ term <fct> 36 months, 60 months, 36 months...  
## $ int\_rate <dbl> 10.65, 15.27, 15.96, 13.49, 12.69,...  
## $ installment <dbl> 162.87, 59.83, 84.33, 339.31, 67.7...  
## $ home\_ownership <fct> RENT, RENT, RENT, RENT, RENT, RENT...  
## $ verification\_status <fct> Verified, Source Verified, Not Ver...  
## $ issue\_d <date> 2011-12-01, 2011-12-01, 2011-12-0...  
## $ purpose <fct> credit\_card, car, small\_business, ...  
## $ dti <dbl> 27.65, 1.00, 8.72, 20.00, 17.94, 1...  
## $ earliest\_cr\_line <date> 1985-01-01, 1999-04-01, 2001-11-0...  
## $ revol\_bal <dbl> 13648, 1687, 2956, 5598, 27783, 79...  
## $ initial\_list\_status <fct> f, f, f, f, f, f, f, f, f, f, f, f...  
## $ recoveries <dbl> 0.00, 117.08, 0.00, 0.00, 0.00, 0....  
## $ collection\_recovery\_fee <dbl> 0.0000, 1.1100, 0.0000, 0.0000, 0....  
## $ application\_type <fct> INDIVIDUAL, INDIVIDUAL, INDIVIDUAL...  
## $ verification\_status\_joint <fct> , , , , , , , , , , , , , , , , , ...  
## $ issue\_year <fct> 2011, 2011, 2011, 2011, 2011, 2011...  
## $ Region <fct> SOUTH\_WEST, SOUTH\_EAST, MID\_WEST, ...  
## $ Employment\_Length <dbl> 10.0, 0.5, 10.0, 10.0, 1.0, 3.0, 8...  
## $ earliest\_cr\_line\_year <fct> 1985, 1999, 2001, 1996, 1996, 2004...

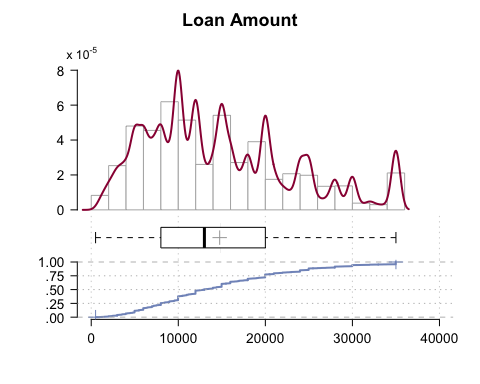
Desc(loan\_df$purpose, main = "Loan purposes", plotit = TRUE)

## -------------------------------------------------------------------------   
## Loan purposes  
##   
## length n NAs unique levels dupes  
## 887'379 887'379 0 14 14 y  
## 100.0% 0.0%   
##   
## level freq perc cumfreq cumperc  
## 1 debt\_consolidation 524'215 59.1% 524'215 59.1%  
## 2 credit\_card 206'182 23.2% 730'397 82.3%  
## 3 home\_improvement 51'829 5.8% 782'226 88.2%  
## 4 other 42'894 4.8% 825'120 93.0%  
## 5 major\_purchase 17'277 1.9% 842'397 94.9%  
## 6 small\_business 10'377 1.2% 852'774 96.1%  
## 7 car 8'863 1.0% 861'637 97.1%  
## 8 medical 8'540 1.0% 870'177 98.1%  
## 9 moving 5'414 0.6% 875'591 98.7%  
## 10 vacation 4'736 0.5% 880'327 99.2%  
## 11 house 3'707 0.4% 884'034 99.6%  
## 12 wedding 2'347 0.3% 886'381 99.9%  
## ... etc.  
## [list output truncated]



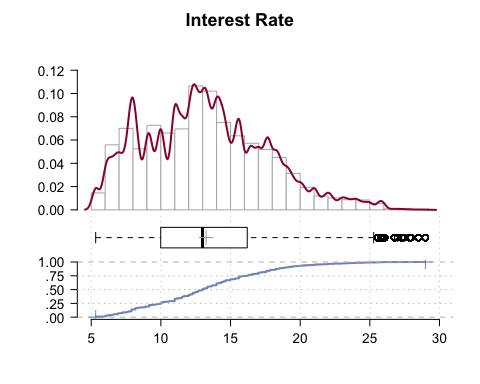
Desc(loan\_df$loan\_amnt, main = "Loan Amount", plotit = TRUE)

## -------------------------------------------------------------------------   
## Loan Amount  
##   
## length n NAs unique 0s mean  
## 887'379 887'379 0 1'372 0 14'755.26  
## 100.0% 0.0% 0.0%   
##   
## .05 .10 .25 median .75 .90  
## 3'600.00 5'000.00 8'000.00 13'000.00 20'000.00 28'000.00  
##   
## range sd vcoef mad IQR skew  
## 34'500.00 8'435.46 0.57 8'599.08 12'000.00 0.68  
##   
## meanCI  
## 14'737.71  
## 14'772.82  
##   
## .95  
## 32'000.00  
##   
## kurt  
## -0.26  
##   
## lowest : 500.0 (11), 550.0, 600.0 (6), 700.0 (3), 725.0  
## highest: 34'900.0 (14), 34'925.0 (9), 34'950.0 (18), 34'975.0 (31), 35'000.0 (36'368)



Desc(loan\_df$int\_rate, main = "Interest Rate", plotit = TRUE)

## -------------------------------------------------------------------------   
## Interest Rate  
##   
## length n NAs unique 0s mean meanCI  
## 887'379 887'379 0 542 0 13.25 13.24  
## 100.0% 0.0% 0.0% 13.26  
##   
## .05 .10 .25 median .75 .90 .95  
## 6.62 7.69 9.99 12.99 16.20 18.99 20.99  
##   
## range sd vcoef mad IQR skew kurt  
## 23.67 4.38 0.33 4.45 6.21 0.43 -0.16  
##   
## lowest : 5.32 (9'651), 5.42 (573), 5.79 (410), 5.93 (1'812), 5.99 (347)  
## highest: 27.49 (7), 27.88 (222), 27.99 (5), 28.49 (139), 28.99 (112)



##check the missing value in each column  
length(loan\_df[,which(is.na(loan\_df$int\_rate))])

## [1] 0

##impute the misssing value with medium for all numeric variables  
for (i in colnames(loan\_df) ){  
 if (class(loan\_df[,i])=="numeric"){  
 loan\_df[is.na(loan\_df[,i]),i] = median(loan\_df[,i], na.rm = TRUE)  
 print (i)  
 print(length(loan\_df[,which(is.na(loan\_df[,i]))]))  
 }  
  
}

## [1] "mths\_since\_last\_major\_derog"  
## [1] 0  
## [1] "mths\_since\_last\_delinq"  
## [1] 0  
## [1] "tot\_coll\_amt"  
## [1] 0  
## [1] "tot\_cur\_bal"  
## [1] 0  
## [1] "total\_rev\_hi\_lim"  
## [1] 0  
## [1] "revol\_util"  
## [1] 0  
## [1] "collections\_12\_mths\_ex\_med"  
## [1] 0  
## [1] "delinq\_2yrs"  
## [1] 0  
## [1] "inq\_last\_6mths"  
## [1] 0  
## [1] "open\_acc"  
## [1] 0  
## [1] "pub\_rec"  
## [1] 0  
## [1] "total\_acc"  
## [1] 0  
## [1] "acc\_now\_delinq"  
## [1] 0  
## [1] "annual\_inc"  
## [1] 0  
## [1] "loan\_amnt"  
## [1] 0  
## [1] "int\_rate"  
## [1] 0  
## [1] "installment"  
## [1] 0  
## [1] "dti"  
## [1] 0  
## [1] "revol\_bal"  
## [1] 0  
## [1] "recoveries"  
## [1] 0  
## [1] "collection\_recovery\_fee"  
## [1] 0  
## [1] "Employment\_Length"  
## [1] 0

summary(loan\_df)

## mths\_since\_last\_major\_derog mths\_since\_last\_delinq tot\_coll\_amt   
## Min. : 0.00 Min. : 0.0 Min. : 0   
## 1st Qu.: 44.00 1st Qu.: 31.0 1st Qu.: 0   
## Median : 44.00 Median : 31.0 Median : 0   
## Mean : 44.03 Mean : 32.5 Mean : 208   
## 3rd Qu.: 44.00 3rd Qu.: 31.0 3rd Qu.: 0   
## Max. :188.00 Max. :188.0 Max. :9152545   
##   
## tot\_cur\_bal total\_rev\_hi\_lim revol\_util   
## Min. : 0 Min. : 0 Min. : 0.00   
## 1st Qu.: 32246 1st Qu.: 14700 1st Qu.: 37.70   
## Median : 80559 Median : 23700 Median : 56.00   
## Mean : 134794 Mean : 31406 Mean : 55.07   
## 3rd Qu.: 195794 3rd Qu.: 37800 3rd Qu.: 73.60   
## Max. :8000078 Max. :9999999 Max. :892.30   
##   
## collections\_12\_mths\_ex\_med delinq\_2yrs inq\_last\_6mths   
## Min. : 0.00000 Min. : 0.0000 Min. : 0.0000   
## 1st Qu.: 0.00000 1st Qu.: 0.0000 1st Qu.: 0.0000   
## Median : 0.00000 Median : 0.0000 Median : 0.0000   
## Mean : 0.01438 Mean : 0.3144 Mean : 0.6946   
## 3rd Qu.: 0.00000 3rd Qu.: 0.0000 3rd Qu.: 1.0000   
## Max. :20.00000 Max. :39.0000 Max. :33.0000   
##   
## open\_acc pub\_rec total\_acc acc\_now\_delinq   
## Min. : 0.00 Min. : 0.0000 Min. : 1.00 Min. : 0.000000   
## 1st Qu.: 8.00 1st Qu.: 0.0000 1st Qu.: 17.00 1st Qu.: 0.000000   
## Median :11.00 Median : 0.0000 Median : 24.00 Median : 0.000000   
## Mean :11.55 Mean : 0.1953 Mean : 25.27 Mean : 0.004991   
## 3rd Qu.:14.00 3rd Qu.: 0.0000 3rd Qu.: 32.00 3rd Qu.: 0.000000   
## Max. :90.00 Max. :86.0000 Max. :169.00 Max. :14.000000   
##   
## annual\_inc loan\_amnt term int\_rate   
## Min. : 0 Min. : 500 36 months:621125 Min. : 5.32   
## 1st Qu.: 45000 1st Qu.: 8000 60 months:266254 1st Qu.: 9.99   
## Median : 65000 Median :13000 Median :12.99   
## Mean : 75028 Mean :14755 Mean :13.25   
## 3rd Qu.: 90000 3rd Qu.:20000 3rd Qu.:16.20   
## Max. :9500000 Max. :35000 Max. :28.99   
##   
## installment home\_ownership verification\_status  
## Min. : 15.67 ANY : 3 Not Verified :266750   
## 1st Qu.: 260.70 MORTGAGE:443557 Source Verified:329558   
## Median : 382.55 NONE : 50 Verified :291071   
## Mean : 436.72 OTHER : 182   
## 3rd Qu.: 572.60 OWN : 87470   
## Max. :1445.46 RENT :356117   
##   
## issue\_d purpose dti   
## Min. :2007-06-01 debt\_consolidation:524215 Min. : 0.00   
## 1st Qu.:2013-12-01 credit\_card :206182 1st Qu.: 11.91   
## Median :2014-11-01 home\_improvement : 51829 Median : 17.65   
## Mean :2014-07-10 other : 42894 Mean : 18.16   
## 3rd Qu.:2015-07-01 major\_purchase : 17277 3rd Qu.: 23.95   
## Max. :2015-12-01 small\_business : 10377 Max. :9999.00   
## (Other) : 34605   
## earliest\_cr\_line revol\_bal initial\_list\_status  
## Min. :1944-01-01 Min. : 0 f:456848   
## 1st Qu.:1994-06-01 1st Qu.: 6443 w:430531   
## Median :1999-09-01 Median : 11875   
## Mean :1998-03-06 Mean : 16921   
## 3rd Qu.:2003-05-01 3rd Qu.: 20829   
## Max. :2012-11-01 Max. :2904836   
## NA's :29   
## recoveries collection\_recovery\_fee application\_type   
## Min. : 0.00 Min. : 0.000 INDIVIDUAL:886868   
## 1st Qu.: 0.00 1st Qu.: 0.000 JOINT : 511   
## Median : 0.00 Median : 0.000   
## Mean : 45.92 Mean : 4.881   
## 3rd Qu.: 0.00 3rd Qu.: 0.000   
## Max. :33520.27 Max. :7002.190   
##   
## verification\_status\_joint issue\_year Region   
## :886868 2015 :421094 MID\_WEST :155029   
## Not Verified : 283 2014 :235628 NORTH\_EAST:204399   
## Source Verified: 61 2013 :134755 SOUTH\_EAST:214646   
## Verified : 167 2012 : 53367 SOUTH\_WEST:104574   
## 2011 : 21721 WEST :208731   
## 2010 : 12537   
## (Other): 8277   
## Employment\_Length earliest\_cr\_line\_year  
## Min. : 0.000 2001 : 63116   
## 1st Qu.: 2.000 2000 : 62614   
## Median : 6.000 2002 : 57410   
## Mean : 5.745 1999 : 56747   
## 3rd Qu.:10.000 2003 : 52730   
## Max. :10.000 (Other):594733   
## NA's : 29

model.df = within(loan\_df, rm("issue\_d", "earliest\_cr\_line"))  
pred.model = lm(int\_rate ~. , model.df)  
summary(pred.model)

##   
## Call:  
## lm(formula = int\_rate ~ ., data = model.df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -66.053 -1.878 -0.251 1.678 90.131   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error t value  
## (Intercept) 5.865e+00 2.451e+00 2.393  
## mths\_since\_last\_major\_derog 4.519e-03 2.950e-04 15.319  
## mths\_since\_last\_delinq 1.606e-03 2.307e-04 6.962  
## tot\_coll\_amt 1.049e-06 2.878e-07 3.646  
## tot\_cur\_bal -1.274e-06 2.615e-08 -48.736  
## total\_rev\_hi\_lim -1.099e-05 1.534e-07 -71.599  
## revol\_util 3.154e-02 1.470e-04 214.581  
## collections\_12\_mths\_ex\_med 4.508e-01 2.139e-02 21.076  
## delinq\_2yrs 2.795e-01 3.895e-03 71.755  
## inq\_last\_6mths 6.655e-01 3.006e-03 221.388  
## open\_acc 4.089e-02 7.898e-04 51.779  
## pub\_rec 3.586e-01 5.076e-03 70.642  
## total\_acc -1.586e-02 3.688e-04 -43.001  
## acc\_now\_delinq 1.077e+00 3.718e-02 28.957  
## annual\_inc -3.102e-06 5.086e-08 -60.993  
## loan\_amnt -1.209e-03 2.136e-06 -566.007  
## term 60 months 1.030e+01 1.276e-02 807.416  
## installment 3.959e-02 6.714e-05 589.644  
## home\_ownershipMORTGAGE -5.572e-01 1.548e+00 -0.360  
## home\_ownershipNONE 1.327e+00 1.598e+00 0.830  
## home\_ownershipOTHER 2.893e-01 1.561e+00 0.185  
## home\_ownershipOWN -1.752e-01 1.548e+00 -0.113  
## home\_ownershipRENT -2.149e-01 1.548e+00 -0.139  
## verification\_statusSource Verified 2.583e-01 7.309e-03 35.340  
## verification\_statusVerified 9.548e-01 7.714e-03 123.766  
## purposecredit\_card -6.704e-01 2.936e-02 -22.836  
## purposedebt\_consolidation 3.520e-01 2.895e-02 12.161  
## purposeeducational 9.373e-01 1.354e-01 6.925  
## purposehome\_improvement 6.141e-01 3.101e-02 19.803  
## purposehouse 2.245e+00 5.255e-02 42.714  
## purposemajor\_purchase 5.381e-01 3.507e-02 15.345  
## purposemedical 2.163e+00 4.072e-02 53.123  
## purposemoving 3.098e+00 4.635e-02 66.835  
## purposeother 2.170e+00 3.136e-02 69.189  
## purposerenewable\_energy 2.581e+00 1.154e-01 22.360  
## purposesmall\_business 2.351e+00 3.894e-02 60.367  
## purposevacation 2.556e+00 4.834e-02 52.867  
## purposewedding 1.586e+00 6.247e-02 25.388  
## dti 6.804e-03 1.714e-04 39.692  
## revol\_bal 9.594e-06 2.504e-07 38.310  
## initial\_list\_statusw -6.298e-01 6.299e-03 -99.991  
## recoveries 2.897e-04 1.177e-05 24.614  
## collection\_recovery\_fee -5.017e-04 7.583e-05 -6.617  
## application\_typeJOINT 1.061e+00 2.077e-01 5.108  
## verification\_status\_jointNot Verified -1.141e+00 2.620e-01 -4.356  
## verification\_status\_jointSource Verified 5.141e-01 4.013e-01 1.281  
## verification\_status\_jointVerified NA NA NA  
## issue\_year2008 9.740e-01 1.249e-01 7.799  
## issue\_year2009 8.651e-01 1.183e-01 7.316  
## issue\_year2010 -3.598e-01 1.151e-01 -3.127  
## issue\_year2011 -3.786e-01 1.142e-01 -3.316  
## issue\_year2012 1.086e+00 1.133e-01 9.585  
## issue\_year2013 1.798e+00 1.131e-01 15.907  
## issue\_year2014 1.395e+00 1.130e-01 12.346  
## issue\_year2015 6.901e-01 1.130e-01 6.107  
## RegionNORTH\_EAST 1.956e-02 9.165e-03 2.134  
## RegionSOUTH\_EAST 1.768e-01 8.973e-03 19.700  
## RegionSOUTH\_WEST 1.254e-01 1.078e-02 11.628  
## RegionWEST 7.264e-02 9.187e-03 7.906  
## Employment\_Length 3.531e-03 7.933e-04 4.450  
## earliest\_cr\_line\_year1946 1.752e+00 2.682e+00 0.653  
## earliest\_cr\_line\_year1948 1.810e+00 3.285e+00 0.551  
## earliest\_cr\_line\_year1949 1.112e-01 3.285e+00 0.034  
## earliest\_cr\_line\_year1950 1.617e+00 2.120e+00 0.763  
## earliest\_cr\_line\_year1951 3.459e+00 2.190e+00 1.580  
## earliest\_cr\_line\_year1952 1.902e+00 2.448e+00 0.777  
## earliest\_cr\_line\_year1953 -8.192e-01 2.244e+00 -0.365  
## earliest\_cr\_line\_year1954 3.259e+00 2.244e+00 1.453  
## earliest\_cr\_line\_year1955 1.716e+00 2.027e+00 0.847  
## earliest\_cr\_line\_year1956 1.496e+00 2.011e+00 0.744  
## earliest\_cr\_line\_year1957 2.963e+00 2.019e+00 1.468  
## earliest\_cr\_line\_year1958 1.417e+00 1.963e+00 0.722  
## earliest\_cr\_line\_year1959 6.562e-01 1.935e+00 0.339  
## earliest\_cr\_line\_year1960 8.426e-01 1.924e+00 0.438  
## earliest\_cr\_line\_year1961 7.154e-01 1.928e+00 0.371  
## earliest\_cr\_line\_year1962 6.669e-01 1.918e+00 0.348  
## earliest\_cr\_line\_year1963 7.624e-01 1.909e+00 0.399  
## earliest\_cr\_line\_year1964 1.018e+00 1.906e+00 0.534  
## earliest\_cr\_line\_year1965 1.167e+00 1.903e+00 0.613  
## earliest\_cr\_line\_year1966 1.054e+00 1.902e+00 0.554  
## earliest\_cr\_line\_year1967 9.169e-01 1.900e+00 0.482  
## earliest\_cr\_line\_year1968 1.041e+00 1.900e+00 0.548  
## earliest\_cr\_line\_year1969 7.325e-01 1.899e+00 0.386  
## earliest\_cr\_line\_year1970 9.588e-01 1.899e+00 0.505  
## earliest\_cr\_line\_year1971 7.509e-01 1.899e+00 0.395  
## earliest\_cr\_line\_year1972 7.667e-01 1.898e+00 0.404  
## earliest\_cr\_line\_year1973 7.855e-01 1.898e+00 0.414  
## earliest\_cr\_line\_year1974 8.946e-01 1.898e+00 0.471  
## earliest\_cr\_line\_year1975 9.762e-01 1.897e+00 0.514  
## earliest\_cr\_line\_year1976 9.959e-01 1.897e+00 0.525  
## earliest\_cr\_line\_year1977 9.506e-01 1.897e+00 0.501  
## earliest\_cr\_line\_year1978 8.972e-01 1.897e+00 0.473  
## earliest\_cr\_line\_year1979 9.790e-01 1.897e+00 0.516  
## earliest\_cr\_line\_year1980 9.614e-01 1.897e+00 0.507  
## earliest\_cr\_line\_year1981 9.902e-01 1.897e+00 0.522  
## earliest\_cr\_line\_year1982 9.255e-01 1.897e+00 0.488  
## earliest\_cr\_line\_year1983 9.571e-01 1.897e+00 0.505  
## earliest\_cr\_line\_year1984 1.068e+00 1.897e+00 0.563  
## earliest\_cr\_line\_year1985 1.046e+00 1.897e+00 0.552  
## earliest\_cr\_line\_year1986 1.079e+00 1.897e+00 0.569  
## earliest\_cr\_line\_year1987 1.108e+00 1.896e+00 0.584  
## earliest\_cr\_line\_year1988 1.111e+00 1.896e+00 0.586  
## earliest\_cr\_line\_year1989 1.115e+00 1.896e+00 0.588  
## earliest\_cr\_line\_year1990 1.117e+00 1.896e+00 0.589  
## earliest\_cr\_line\_year1991 1.150e+00 1.896e+00 0.606  
## earliest\_cr\_line\_year1992 1.162e+00 1.896e+00 0.613  
## earliest\_cr\_line\_year1993 1.196e+00 1.896e+00 0.630  
## earliest\_cr\_line\_year1994 1.233e+00 1.896e+00 0.650  
## earliest\_cr\_line\_year1995 1.254e+00 1.896e+00 0.661  
## earliest\_cr\_line\_year1996 1.288e+00 1.896e+00 0.679  
## earliest\_cr\_line\_year1997 1.293e+00 1.896e+00 0.682  
## earliest\_cr\_line\_year1998 1.297e+00 1.896e+00 0.684  
## earliest\_cr\_line\_year1999 1.323e+00 1.896e+00 0.698  
## earliest\_cr\_line\_year2000 1.350e+00 1.896e+00 0.712  
## earliest\_cr\_line\_year2001 1.407e+00 1.896e+00 0.742  
## earliest\_cr\_line\_year2002 1.485e+00 1.896e+00 0.783  
## earliest\_cr\_line\_year2003 1.558e+00 1.896e+00 0.822  
## earliest\_cr\_line\_year2004 1.705e+00 1.896e+00 0.899  
## earliest\_cr\_line\_year2005 1.806e+00 1.896e+00 0.953  
## earliest\_cr\_line\_year2006 1.914e+00 1.896e+00 1.009  
## earliest\_cr\_line\_year2007 2.160e+00 1.896e+00 1.139  
## earliest\_cr\_line\_year2008 2.472e+00 1.896e+00 1.304  
## earliest\_cr\_line\_year2009 2.712e+00 1.897e+00 1.430  
## earliest\_cr\_line\_year2010 2.960e+00 1.897e+00 1.560  
## earliest\_cr\_line\_year2011 3.218e+00 1.897e+00 1.697  
## earliest\_cr\_line\_year2012 3.489e+00 1.897e+00 1.839  
## Pr(>|t|)   
## (Intercept) 0.016720 \*   
## mths\_since\_last\_major\_derog < 2e-16 \*\*\*  
## mths\_since\_last\_delinq 3.35e-12 \*\*\*  
## tot\_coll\_amt 0.000267 \*\*\*  
## tot\_cur\_bal < 2e-16 \*\*\*  
## total\_rev\_hi\_lim < 2e-16 \*\*\*  
## revol\_util < 2e-16 \*\*\*  
## collections\_12\_mths\_ex\_med < 2e-16 \*\*\*  
## delinq\_2yrs < 2e-16 \*\*\*  
## inq\_last\_6mths < 2e-16 \*\*\*  
## open\_acc < 2e-16 \*\*\*  
## pub\_rec < 2e-16 \*\*\*  
## total\_acc < 2e-16 \*\*\*  
## acc\_now\_delinq < 2e-16 \*\*\*  
## annual\_inc < 2e-16 \*\*\*  
## loan\_amnt < 2e-16 \*\*\*  
## term 60 months < 2e-16 \*\*\*  
## installment < 2e-16 \*\*\*  
## home\_ownershipMORTGAGE 0.718990   
## home\_ownershipNONE 0.406522   
## home\_ownershipOTHER 0.852999   
## home\_ownershipOWN 0.909904   
## home\_ownershipRENT 0.889600   
## verification\_statusSource Verified < 2e-16 \*\*\*  
## verification\_statusVerified < 2e-16 \*\*\*  
## purposecredit\_card < 2e-16 \*\*\*  
## purposedebt\_consolidation < 2e-16 \*\*\*  
## purposeeducational 4.37e-12 \*\*\*  
## purposehome\_improvement < 2e-16 \*\*\*  
## purposehouse < 2e-16 \*\*\*  
## purposemajor\_purchase < 2e-16 \*\*\*  
## purposemedical < 2e-16 \*\*\*  
## purposemoving < 2e-16 \*\*\*  
## purposeother < 2e-16 \*\*\*  
## purposerenewable\_energy < 2e-16 \*\*\*  
## purposesmall\_business < 2e-16 \*\*\*  
## purposevacation < 2e-16 \*\*\*  
## purposewedding < 2e-16 \*\*\*  
## dti < 2e-16 \*\*\*  
## revol\_bal < 2e-16 \*\*\*  
## initial\_list\_statusw < 2e-16 \*\*\*  
## recoveries < 2e-16 \*\*\*  
## collection\_recovery\_fee 3.67e-11 \*\*\*  
## application\_typeJOINT 3.26e-07 \*\*\*  
## verification\_status\_jointNot Verified 1.32e-05 \*\*\*  
## verification\_status\_jointSource Verified 0.200127   
## verification\_status\_jointVerified NA   
## issue\_year2008 6.25e-15 \*\*\*  
## issue\_year2009 2.55e-13 \*\*\*  
## issue\_year2010 0.001764 \*\*   
## issue\_year2011 0.000913 \*\*\*  
## issue\_year2012 < 2e-16 \*\*\*  
## issue\_year2013 < 2e-16 \*\*\*  
## issue\_year2014 < 2e-16 \*\*\*  
## issue\_year2015 1.01e-09 \*\*\*  
## RegionNORTH\_EAST 0.032859 \*   
## RegionSOUTH\_EAST < 2e-16 \*\*\*  
## RegionSOUTH\_WEST < 2e-16 \*\*\*  
## RegionWEST 2.65e-15 \*\*\*  
## Employment\_Length 8.58e-06 \*\*\*  
## earliest\_cr\_line\_year1946 0.513666   
## earliest\_cr\_line\_year1948 0.581528   
## earliest\_cr\_line\_year1949 0.972989   
## earliest\_cr\_line\_year1950 0.445604   
## earliest\_cr\_line\_year1951 0.114213   
## earliest\_cr\_line\_year1952 0.437224   
## earliest\_cr\_line\_year1953 0.715052   
## earliest\_cr\_line\_year1954 0.146351   
## earliest\_cr\_line\_year1955 0.397209   
## earliest\_cr\_line\_year1956 0.457171   
## earliest\_cr\_line\_year1957 0.142193   
## earliest\_cr\_line\_year1958 0.470342   
## earliest\_cr\_line\_year1959 0.734596   
## earliest\_cr\_line\_year1960 0.661510   
## earliest\_cr\_line\_year1961 0.710553   
## earliest\_cr\_line\_year1962 0.728005   
## earliest\_cr\_line\_year1963 0.689576   
## earliest\_cr\_line\_year1964 0.593131   
## earliest\_cr\_line\_year1965 0.539892   
## earliest\_cr\_line\_year1966 0.579586   
## earliest\_cr\_line\_year1967 0.629454   
## earliest\_cr\_line\_year1968 0.583965   
## earliest\_cr\_line\_year1969 0.699730   
## earliest\_cr\_line\_year1970 0.613611   
## earliest\_cr\_line\_year1971 0.692478   
## earliest\_cr\_line\_year1972 0.686242   
## earliest\_cr\_line\_year1973 0.678954   
## earliest\_cr\_line\_year1974 0.637333   
## earliest\_cr\_line\_year1975 0.606928   
## earliest\_cr\_line\_year1976 0.599617   
## earliest\_cr\_line\_year1977 0.616281   
## earliest\_cr\_line\_year1978 0.636214   
## earliest\_cr\_line\_year1979 0.605792   
## earliest\_cr\_line\_year1980 0.612276   
## earliest\_cr\_line\_year1981 0.601646   
## earliest\_cr\_line\_year1982 0.625571   
## earliest\_cr\_line\_year1983 0.613830   
## earliest\_cr\_line\_year1984 0.573239   
## earliest\_cr\_line\_year1985 0.581093   
## earliest\_cr\_line\_year1986 0.569552   
## earliest\_cr\_line\_year1987 0.559214   
## earliest\_cr\_line\_year1988 0.557943   
## earliest\_cr\_line\_year1989 0.556610   
## earliest\_cr\_line\_year1990 0.555729   
## earliest\_cr\_line\_year1991 0.544400   
## earliest\_cr\_line\_year1992 0.539931   
## earliest\_cr\_line\_year1993 0.528394   
## earliest\_cr\_line\_year1994 0.515737   
## earliest\_cr\_line\_year1995 0.508569   
## earliest\_cr\_line\_year1996 0.496975   
## earliest\_cr\_line\_year1997 0.495443   
## earliest\_cr\_line\_year1998 0.493888   
## earliest\_cr\_line\_year1999 0.485453   
## earliest\_cr\_line\_year2000 0.476390   
## earliest\_cr\_line\_year2001 0.458120   
## earliest\_cr\_line\_year2002 0.433595   
## earliest\_cr\_line\_year2003 0.411295   
## earliest\_cr\_line\_year2004 0.368658   
## earliest\_cr\_line\_year2005 0.340813   
## earliest\_cr\_line\_year2006 0.312833   
## earliest\_cr\_line\_year2007 0.254796   
## earliest\_cr\_line\_year2008 0.192329   
## earliest\_cr\_line\_year2009 0.152760   
## earliest\_cr\_line\_year2010 0.118654   
## earliest\_cr\_line\_year2011 0.089770 .   
## earliest\_cr\_line\_year2012 0.065905 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.682 on 887225 degrees of freedom  
## (29 observations deleted due to missingness)  
## Multiple R-squared: 0.6255, Adjusted R-squared: 0.6254   
## F-statistic: 1.195e+04 on 124 and 887225 DF, p-value: < 2.2e-16

## earliest\_cr\_line\_year/ home\_ownership/ mths\_since\_last\_delinq is not promising. ##delete this feature  
model.df=within(model.df, rm("earliest\_cr\_line\_year", "home\_ownership", "mths\_since\_last\_delinq"))  
pred.model = lm(int\_rate ~. , model.df)  
summary(pred.model)

##   
## Call:  
## lm(formula = int\_rate ~ ., data = model.df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -66.710 -1.911 -0.253 1.694 100.937   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error t value  
## (Intercept) 7.234e+00 1.158e-01 62.491  
## mths\_since\_last\_major\_derog 3.855e-03 2.762e-04 13.955  
## tot\_coll\_amt 8.503e-07 2.909e-07 2.923  
## tot\_cur\_bal -1.891e-06 2.382e-08 -79.393  
## total\_rev\_hi\_lim -1.188e-05 1.547e-07 -76.755  
## revol\_util 3.061e-02 1.482e-04 206.554  
## collections\_12\_mths\_ex\_med 4.123e-01 2.160e-02 19.083  
## delinq\_2yrs 2.344e-01 3.614e-03 64.855  
## inq\_last\_6mths 6.741e-01 3.036e-03 222.056  
## open\_acc 5.575e-02 7.869e-04 70.851  
## pub\_rec 2.831e-01 5.091e-03 55.607  
## total\_acc -3.040e-02 3.534e-04 -86.013  
## acc\_now\_delinq 1.000e+00 3.747e-02 26.691  
## annual\_inc -3.071e-06 5.125e-08 -59.917  
## loan\_amnt -1.238e-03 2.146e-06 -576.728  
## term 60 months 1.041e+01 1.285e-02 810.188  
## installment 4.035e-02 6.757e-05 597.217  
## verification\_statusSource Verified 2.881e-01 7.381e-03 39.031  
## verification\_statusVerified 9.502e-01 7.785e-03 122.066  
## purposecredit\_card -6.970e-01 2.967e-02 -23.492  
## purposedebt\_consolidation 3.177e-01 2.925e-02 10.861  
## purposeeducational 9.549e-01 1.368e-01 6.981  
## purposehome\_improvement 4.874e-01 3.130e-02 15.574  
## purposehouse 2.287e+00 5.311e-02 43.071  
## purposemajor\_purchase 5.516e-01 3.544e-02 15.561  
## purposemedical 2.093e+00 4.115e-02 50.857  
## purposemoving 3.179e+00 4.681e-02 67.902  
## purposeother 2.159e+00 3.169e-02 68.120  
## purposerenewable\_energy 2.588e+00 1.167e-01 22.183  
## purposesmall\_business 2.333e+00 3.936e-02 59.267  
## purposevacation 2.574e+00 4.886e-02 52.681  
## purposewedding 1.619e+00 6.313e-02 25.653  
## dti 6.809e-03 1.732e-04 39.303  
## revol\_bal 9.505e-06 2.527e-07 37.615  
## initial\_list\_statusw -6.409e-01 6.365e-03 -100.691  
## recoveries 3.017e-04 1.190e-05 25.360  
## collection\_recovery\_fee -5.340e-04 7.660e-05 -6.972  
## application\_typeJOINT 1.022e+00 2.099e-01 4.871  
## verification\_status\_jointNot Verified -1.111e+00 2.648e-01 -4.196  
## verification\_status\_jointSource Verified 5.825e-01 4.056e-01 1.436  
## verification\_status\_jointVerified NA NA NA  
## issue\_year2008 1.035e+00 1.236e-01 8.371  
## issue\_year2009 9.393e-01 1.167e-01 8.047  
## issue\_year2010 -2.544e-01 1.134e-01 -2.242  
## issue\_year2011 -2.411e-01 1.125e-01 -2.143  
## issue\_year2012 1.288e+00 1.116e-01 11.540  
## issue\_year2013 2.043e+00 1.114e-01 18.349  
## issue\_year2014 1.699e+00 1.113e-01 15.273  
## issue\_year2015 1.070e+00 1.113e-01 9.615  
## RegionNORTH\_EAST 9.736e-02 9.182e-03 10.603  
## RegionSOUTH\_EAST 2.023e-01 9.065e-03 22.315  
## RegionSOUTH\_WEST 1.486e-01 1.089e-02 13.638  
## RegionWEST 1.572e-01 9.153e-03 17.176  
## Employment\_Length -1.662e-02 7.830e-04 -21.229  
## Pr(>|t|)   
## (Intercept) < 2e-16 \*\*\*  
## mths\_since\_last\_major\_derog < 2e-16 \*\*\*  
## tot\_coll\_amt 0.00347 \*\*   
## tot\_cur\_bal < 2e-16 \*\*\*  
## total\_rev\_hi\_lim < 2e-16 \*\*\*  
## revol\_util < 2e-16 \*\*\*  
## collections\_12\_mths\_ex\_med < 2e-16 \*\*\*  
## delinq\_2yrs < 2e-16 \*\*\*  
## inq\_last\_6mths < 2e-16 \*\*\*  
## open\_acc < 2e-16 \*\*\*  
## pub\_rec < 2e-16 \*\*\*  
## total\_acc < 2e-16 \*\*\*  
## acc\_now\_delinq < 2e-16 \*\*\*  
## annual\_inc < 2e-16 \*\*\*  
## loan\_amnt < 2e-16 \*\*\*  
## term 60 months < 2e-16 \*\*\*  
## installment < 2e-16 \*\*\*  
## verification\_statusSource Verified < 2e-16 \*\*\*  
## verification\_statusVerified < 2e-16 \*\*\*  
## purposecredit\_card < 2e-16 \*\*\*  
## purposedebt\_consolidation < 2e-16 \*\*\*  
## purposeeducational 2.93e-12 \*\*\*  
## purposehome\_improvement < 2e-16 \*\*\*  
## purposehouse < 2e-16 \*\*\*  
## purposemajor\_purchase < 2e-16 \*\*\*  
## purposemedical < 2e-16 \*\*\*  
## purposemoving < 2e-16 \*\*\*  
## purposeother < 2e-16 \*\*\*  
## purposerenewable\_energy < 2e-16 \*\*\*  
## purposesmall\_business < 2e-16 \*\*\*  
## purposevacation < 2e-16 \*\*\*  
## purposewedding < 2e-16 \*\*\*  
## dti < 2e-16 \*\*\*  
## revol\_bal < 2e-16 \*\*\*  
## initial\_list\_statusw < 2e-16 \*\*\*  
## recoveries < 2e-16 \*\*\*  
## collection\_recovery\_fee 3.14e-12 \*\*\*  
## application\_typeJOINT 1.11e-06 \*\*\*  
## verification\_status\_jointNot Verified 2.72e-05 \*\*\*  
## verification\_status\_jointSource Verified 0.15090   
## verification\_status\_jointVerified NA   
## issue\_year2008 < 2e-16 \*\*\*  
## issue\_year2009 8.49e-16 \*\*\*  
## issue\_year2010 0.02495 \*   
## issue\_year2011 0.03213 \*   
## issue\_year2012 < 2e-16 \*\*\*  
## issue\_year2013 < 2e-16 \*\*\*  
## issue\_year2014 < 2e-16 \*\*\*  
## issue\_year2015 < 2e-16 \*\*\*  
## RegionNORTH\_EAST < 2e-16 \*\*\*  
## RegionSOUTH\_EAST < 2e-16 \*\*\*  
## RegionSOUTH\_WEST < 2e-16 \*\*\*  
## RegionWEST < 2e-16 \*\*\*  
## Employment\_Length < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.711 on 887326 degrees of freedom  
## Multiple R-squared: 0.6173, Adjusted R-squared: 0.6173   
## F-statistic: 2.753e+04 on 52 and 887326 DF, p-value: < 2.2e-16

# in output  
# Multiple R-squared: 0.6176, Adjusted R-squared: 0.6175   
# if R-square <= 0.3 it's pretty poor  
# F-statistic: 2.703e+04 on 53 and 887325 DF, p-value: < 2.2e-16  
# F test => if the variance interpreted by the model vs. not interpreted by the model is significant.   
# more siginificant is better.