5 continuous & 5 categorical variables have strong relationship with response variable

[redacted]

```
loan <- read.csv("loan.csv", stringsAsFactors = FALSE)
num.NA <- sort(sapply(loan, function(x){sum(is.na(x))}),
               decreasing = TRUE)
remain.col <- names(num.NA)[which(num.NA <= 0.8 * dim(loan)[1])]
loan <- loan[, remain.col]
loan$annual_inc[which(is.na(loan$annual_inc))] <-
  median(loan$annual_inc, na.rm = T)
```

find numeric features

```
n <- ncol(loan)
numeric.check <- rep(NA, n)
for(i in 1:n){
  numeric.check[i] <- is.numeric(loan[,i])
}
numerical.v <-which(numeric.check ==TRUE)
#after checking data, find all col in policy_code are same, which will incur error in co
rrelation matrix
numerical.v <- numerical.v[-34]
cor <- cor(loan$int_rate, loan[, numerical.v],
           use = 'pairwise.complete.obs')
name <- names(loan[,numerical.v])
features <-name[order(abs(cor), decreasing = TRUE)[2:6]]
features
```

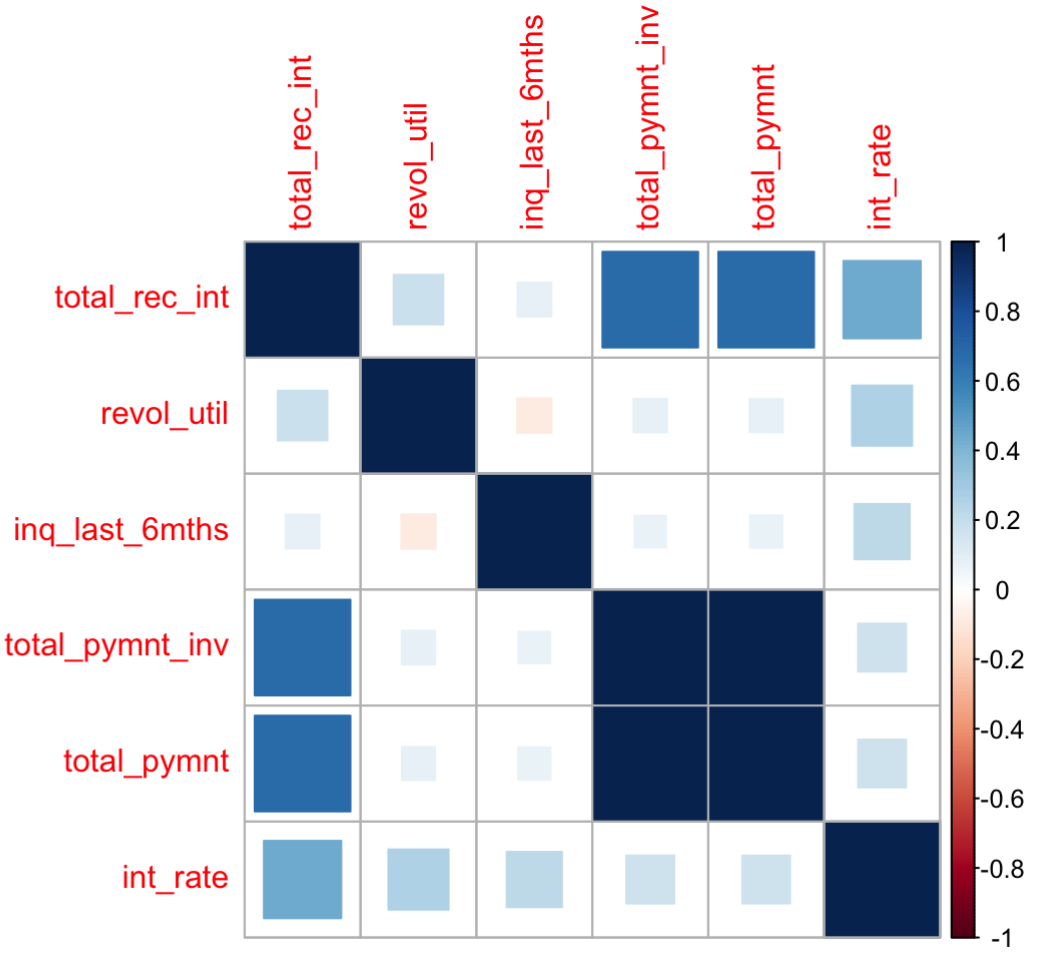
```
## [1] "total_rec_int"   "revol_util"      "inq_last_6mths"  "total_pymnt_inv"
## [5] "total_pymnt"
```

```
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 3.4.2
```

```
## corrplot 0.84 loaded
```

```
cor5 <- cor( loan[,c(features, "int_rate")],
             use = 'pairwise.complete.obs')
corrplot(cor5, method = 'square', tl.cex = 1)
```



find 5

categorical features These are not considered after reviewing the data: {emp_title, loan status, initial_list_status, next_pymnt_d, zip_code, desc, issue_d_1, issue_year, title,last_pymnt_d and url}

```
numerical.c <- which(numeric.check== FALSE)
length(numerical.c) #25 variables
```

```
## [1] 23
```

```
#1"term" [reserve]
table(loan[,numerical.c[1]])
```

```
##
##  36 months  60 months
##     621125     266254
```

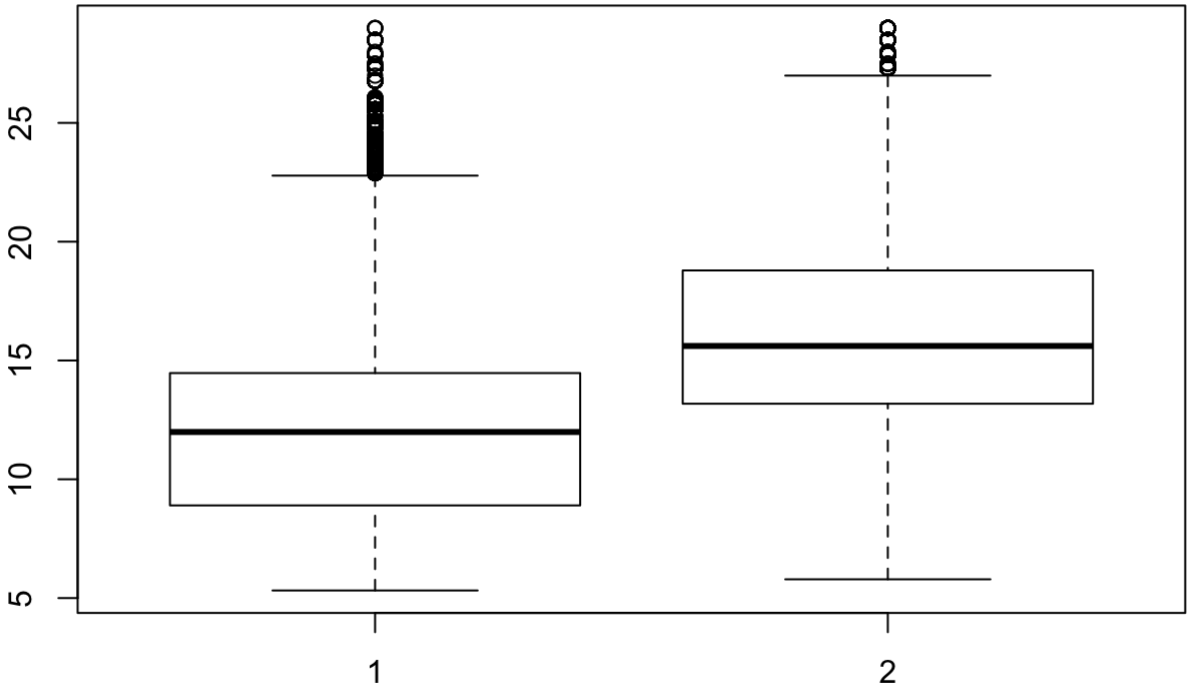
```
boxplot(subset(loan, term ==' 36 months')$int_rate,
        subset(loan, term ==' 60 months')$int_rate)
#2"grade"  [leave subgrade]
table(loan[,numerical.c[2]])
```

```
##
##      A      B      C      D      E      F      G
## 148202 254535 245860 139542  70705  23046   5489
```

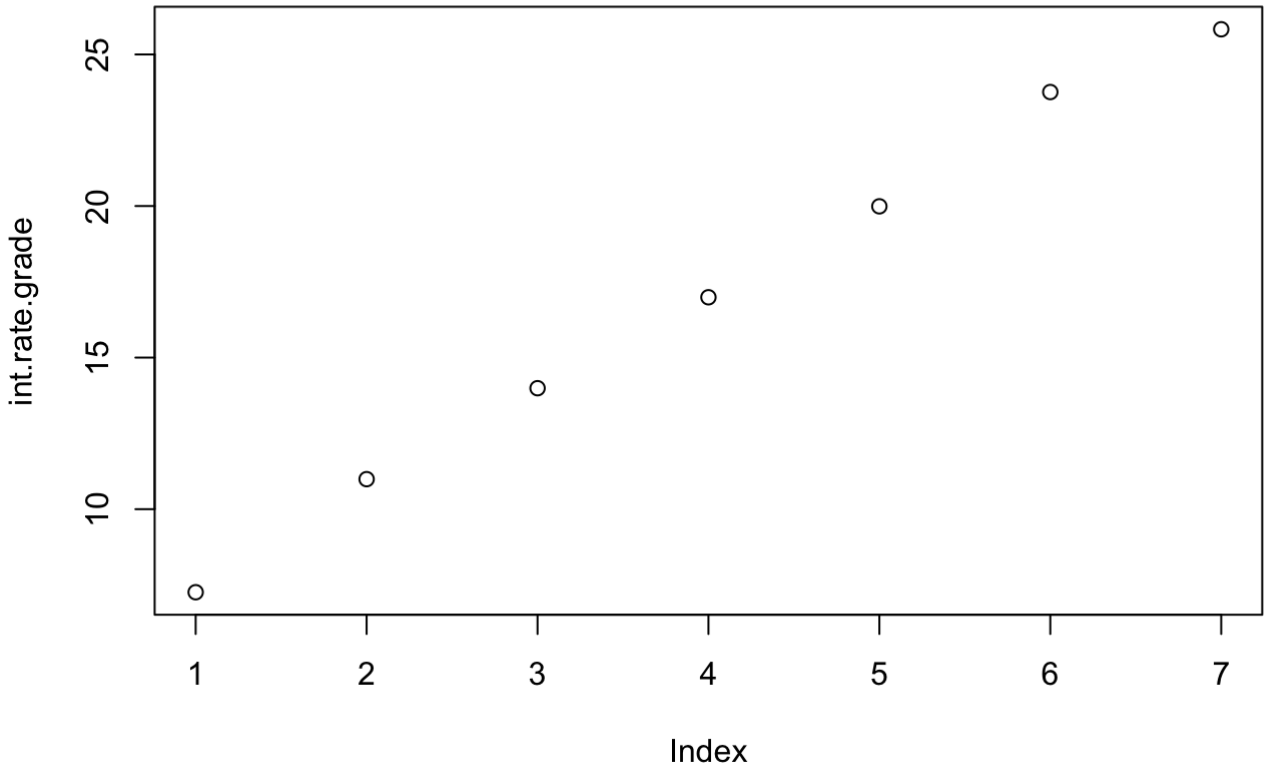
```
library(zoo)
```

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

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



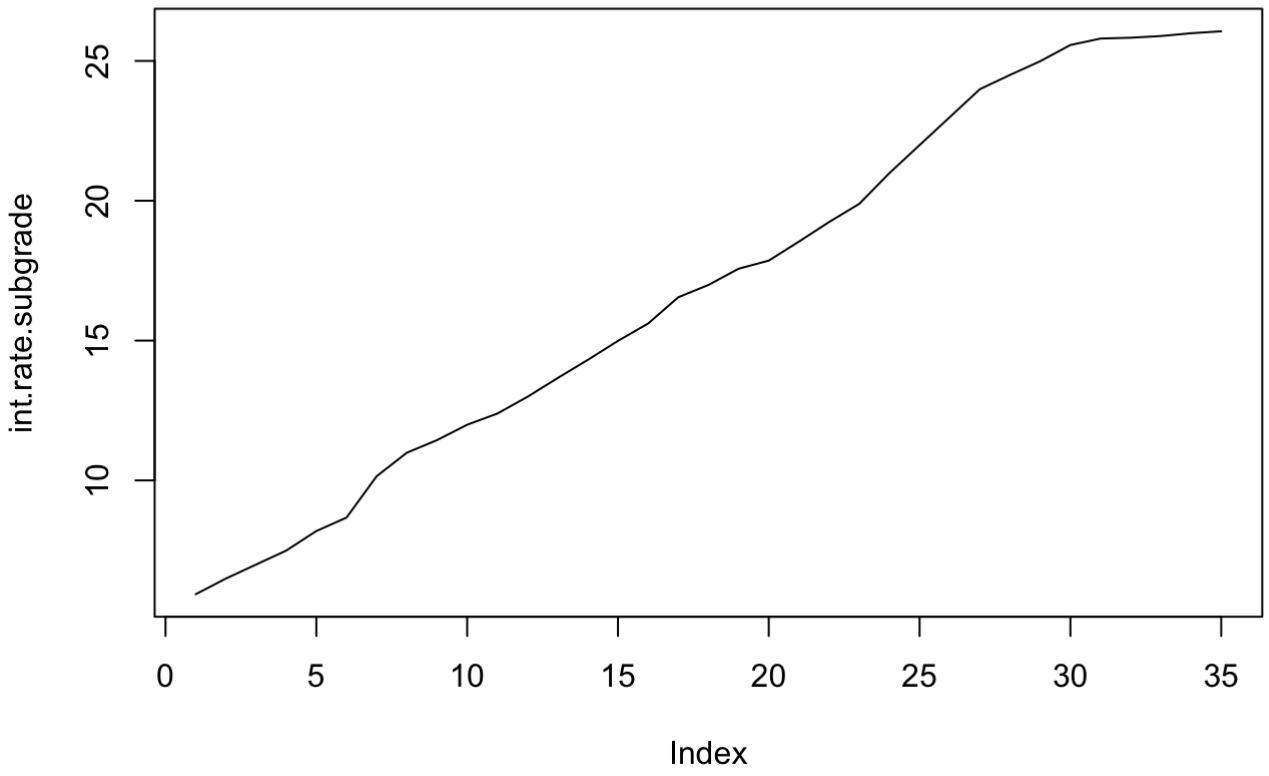

```
int.rate.grade <- by(loan, loan$grade,
                     function(x){return(median(x$int_rate))})
plot(int.rate.grade) #reverse
```




```
#3"sub_grade" [reserve]
table(loan[,numerical.c[3]])
```

```
##
##    A1    A2    A3    A4    A5    B1    B2    B3    B4    B5    C1    C2
## 22913 22485 23457 34531 44816 44972 48781 56323 55626 48833 53387 52236
##    C3    C4    C5    D1    D2    D3    D4    D5    E1    E2    E3    E4
## 50161 48857 41219 36238 29803 26554 25558 21389 18268 17004 14134 11724
##    E5    F1    F2    F3    F4    F5    G1    G2    G3    G4    G5
##  9575  7218  5392  4433  3409  2594  1871  1398   981   663   576
```

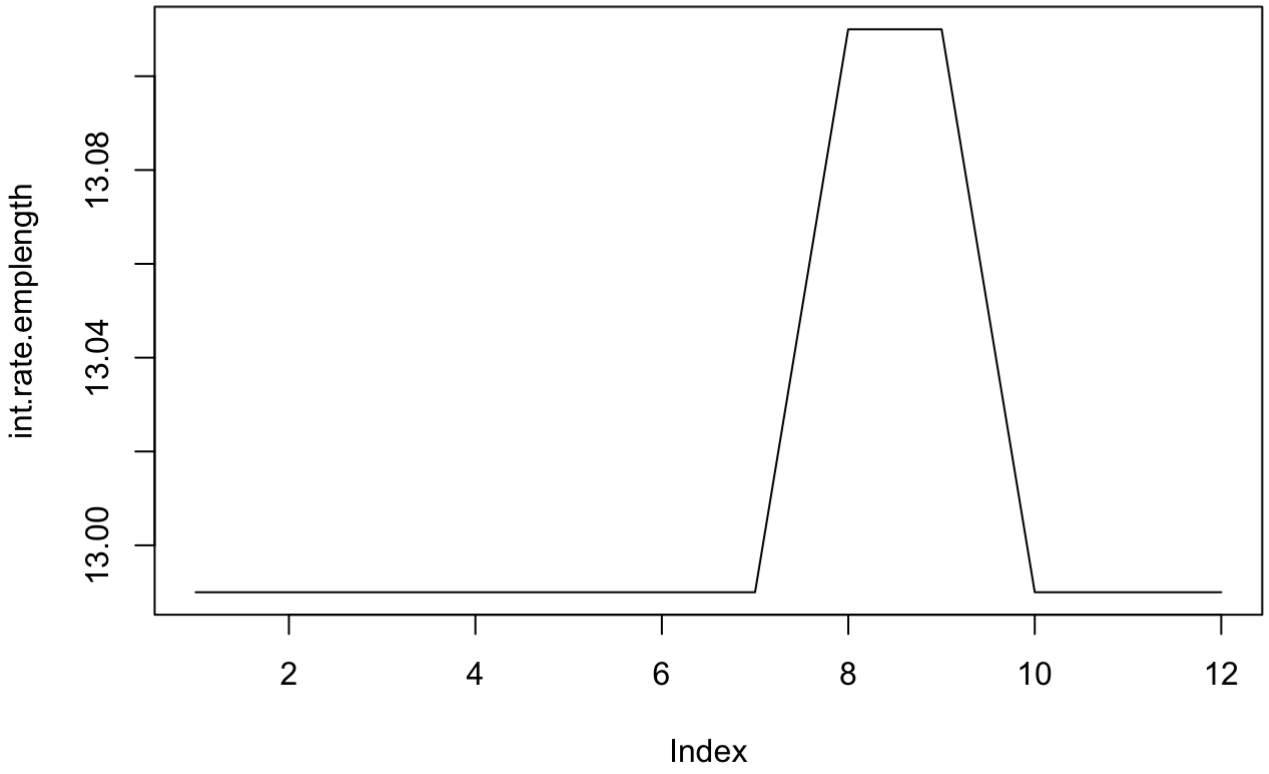
```
int.rate.subgrade <- by(loan, loan$sub_grade,
                    function(x){return(median(x$int_rate))})
plot(int.rate.subgrade, type = 'l') #reverse
```




```
#5"emp_length"[not consider]
table(loan[,numerical.c[5]])
```

```
##
##  < 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
```

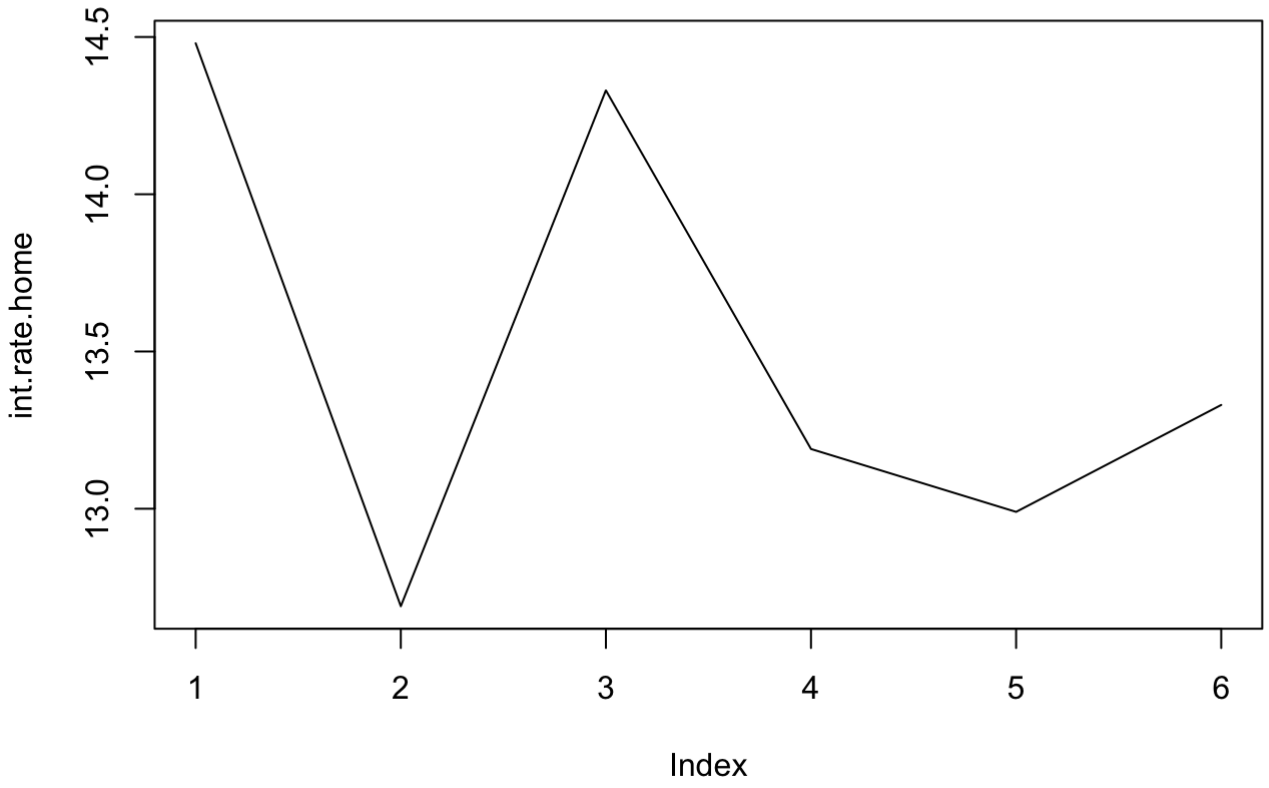
```
int.rate.emplength <- by(loan, loan$emp_length,
                         function(x){return(median(x$int_rate))})
plot(int.rate.emplength, type = 'l') #not considered, not too much inflence
```




```
#6"home_ownership"[reserve]
table(loan[,numerical.c[6]])
```

```
##
##      ANY MORTGAGE     NONE    OTHER      OWN     RENT
##        3   443557       50      182    87470   356117
```

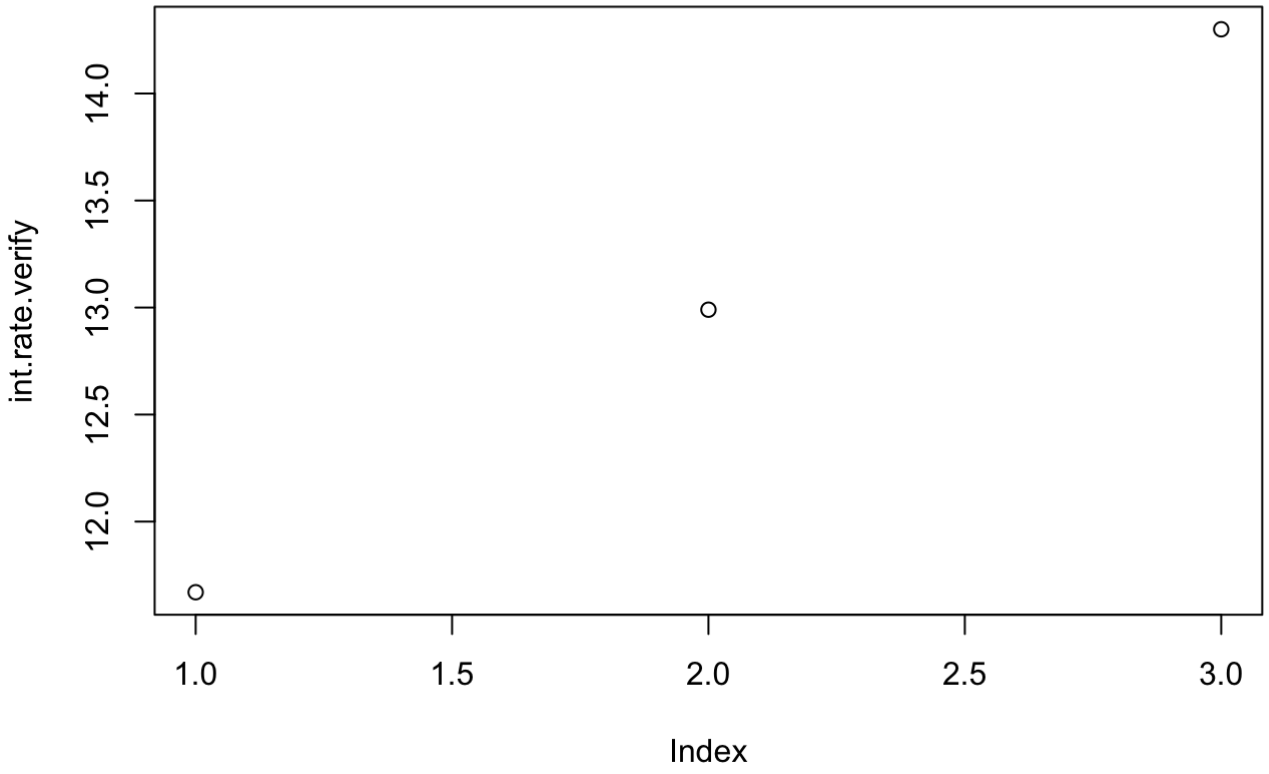
```
int.rate.home <- by(loan, loan$home_ownership,
                        function(x){return(median(x$int_rate))})
plot( int.rate.home, type = 'l')
```




```
#7"verification_status" [reserve]
table(loan[,numerical.c[7]])
```

```
##
##    Not Verified Source Verified        Verified
##          266750          329558          291071
```

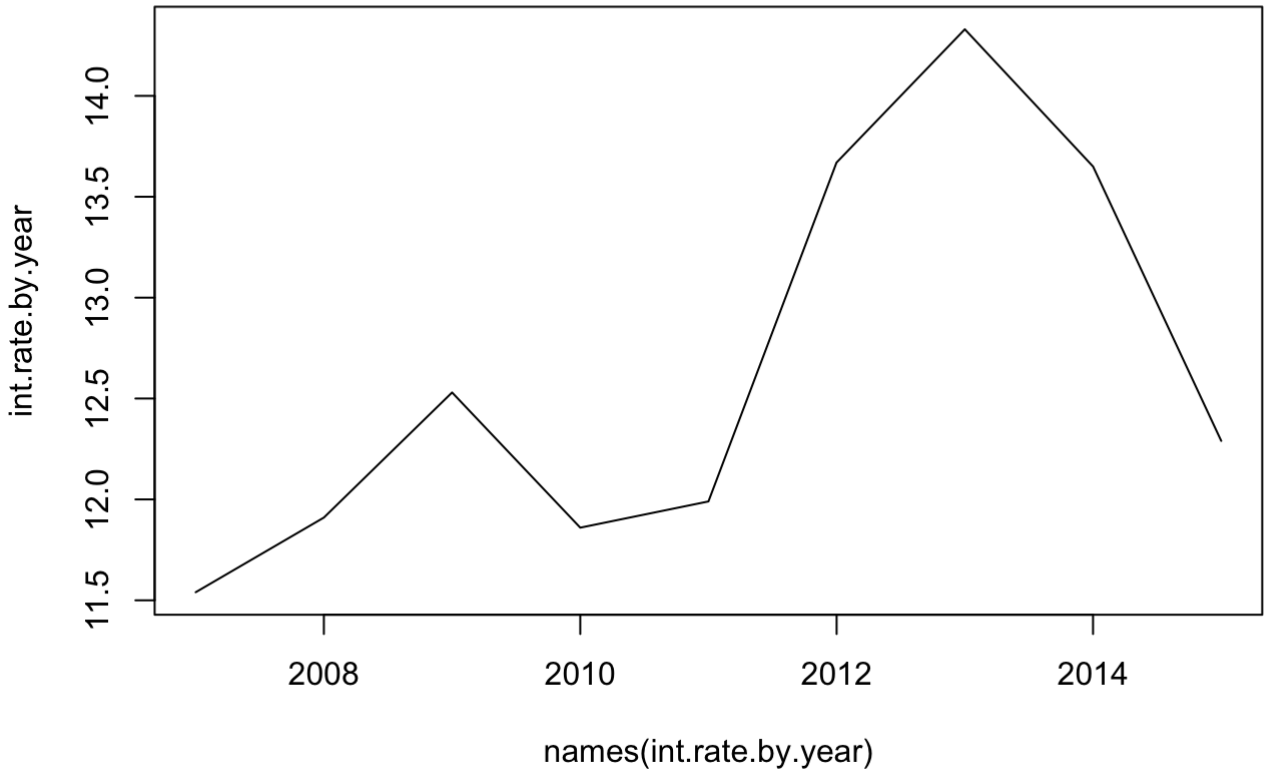
```
int.rate.verify <- by(loan, loan$verification_status,
                      function(x){return(median(x$int_rate))})
plot( int.rate.verify) #not consider
```




```
#8"issue_year" [reserved]
loan$issue_d_1 <-as.Date(as.yearmon(loan$issue_d, '%b-%Y'))
loan$issue_year <- format(loan$issue_d_1, '%Y')
table(loan$issue_year)
```

```
##
##   2007   2008   2009   2010   2011   2012   2013   2014   2015
##    603   2393   5281  12537  21721  53367 134755 235628 421094
```

```
int.rate.by.year <- by(loan, loan$issue_year,
                       function(x){return(median(x$int_rate))})
plot(names(int.rate.by.year), int.rate.by.year, type = 'l')
```



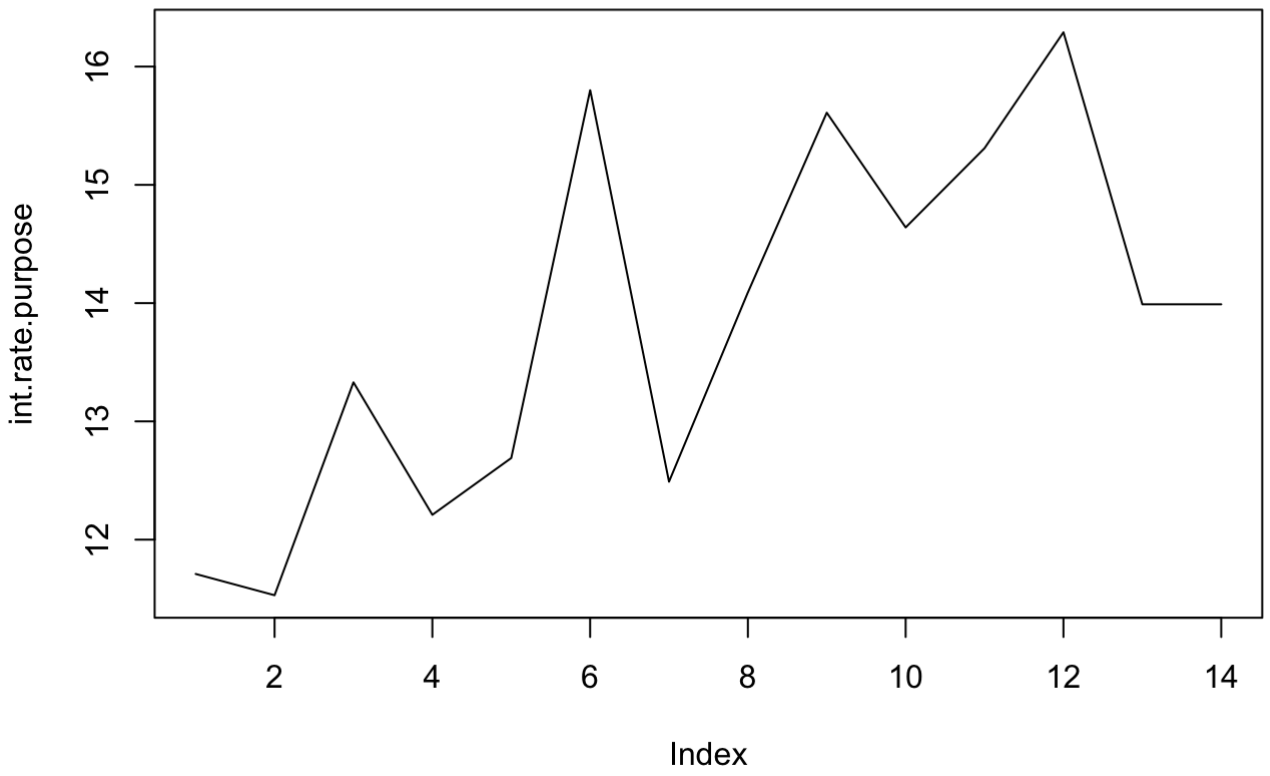

```
#10"pymnt_plan"
table(loan[, "pymnt_plan"]) #not considered
```

```
##
##      n      y
## 887369     10
```

```
#13"purpose" [reserve]
table(loan[, "purpose"])
```

```
##
##                car        credit_card debt_consolidation
##               8863             206182             524215
##        educational   home_improvement              house
##                423              51829               3707
##     major_purchase            medical             moving
##              17277               8540               5414
##              other   renewable_energy     small_business
##              42894                575              10377
##           vacation            wedding
##               4736               2347
```

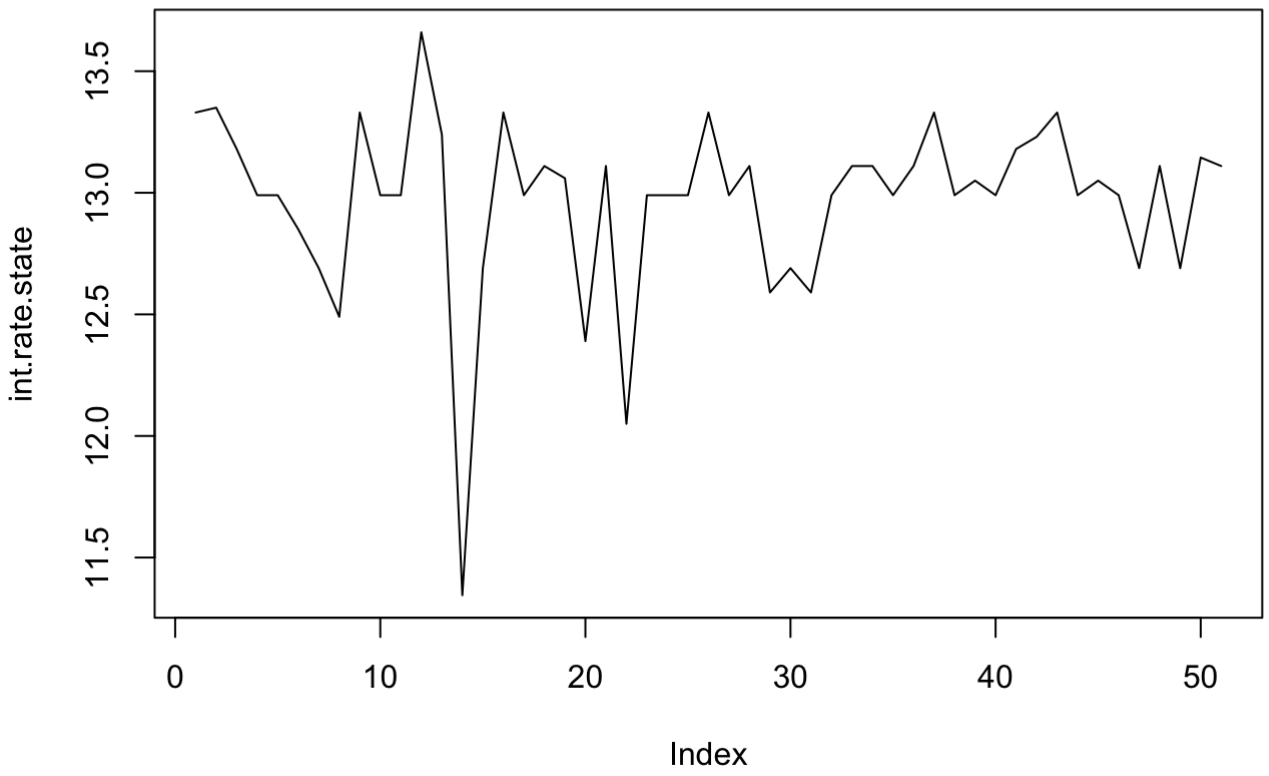
```
int.rate.purpose <- by(loan, loan$purpose,
                       function(x){return(median(x$int_rate))})
plot(int.rate.purpose, type = 'l')
```




```
#16"addr_state"
table(loan[,"addr_state"])
```

```
##
##     AK     AL     AR     AZ     CA     CO     CT     DC     DE     FL
##   2205  11200   6640  20412 129517  18807  13531   2432   2511  60935
##     GA     HI     IA     ID     IL     IN     KS     KY     LA     MA
##  29085   4570     14     12  35476  13789   7926   8550  10587  20593
##     MD     ME     MI     MN     MO     MS     MT     NC     ND     NE
##  21031    525  22985  15957  14207   3819   2558  24720    479   1176
##     NH     NJ     NM     NV     NY     OH     OK     OR     PA     RI
##   4294  33256   4939  12443  74086  29631   8085  10893  31393   3893
##     SC     SD     TN     TX     UT     VA     VT     WA     WI     WV
##  10639   1815  12887  71138   6264  26255   1797  19434  11574   4386
##     WY
##   2028
```

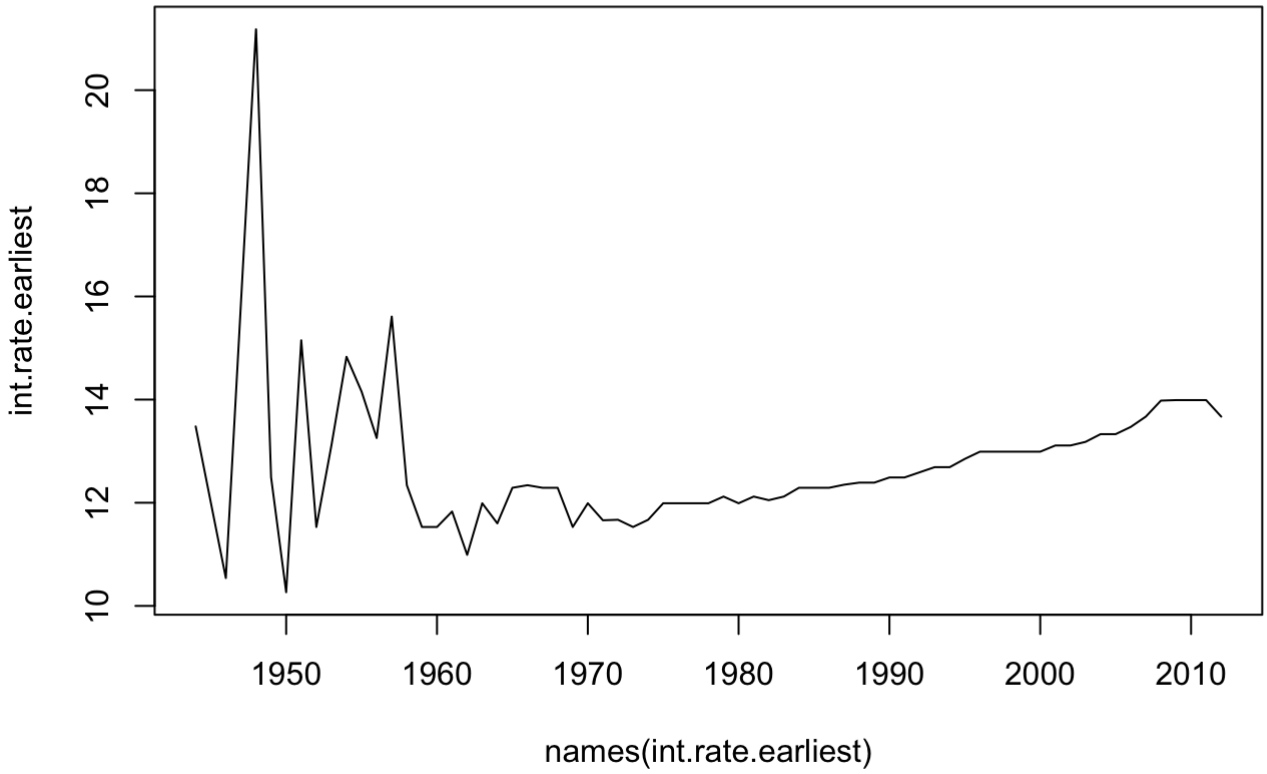
```
int.rate.state <- by(loan, loan$addr_state,
                     function(x){return(median(x$int_rate))})
plot(int.rate.state, type = 'l')#not considered
```




```
#17"earliest_cr_line" [reserve]
head(loan[,"earliest_cr_line"])
```

```
## [1] "Jan-1985" "Apr-1999" "Nov-2001" "Feb-1996" "Jan-1996" "Nov-2004"
```

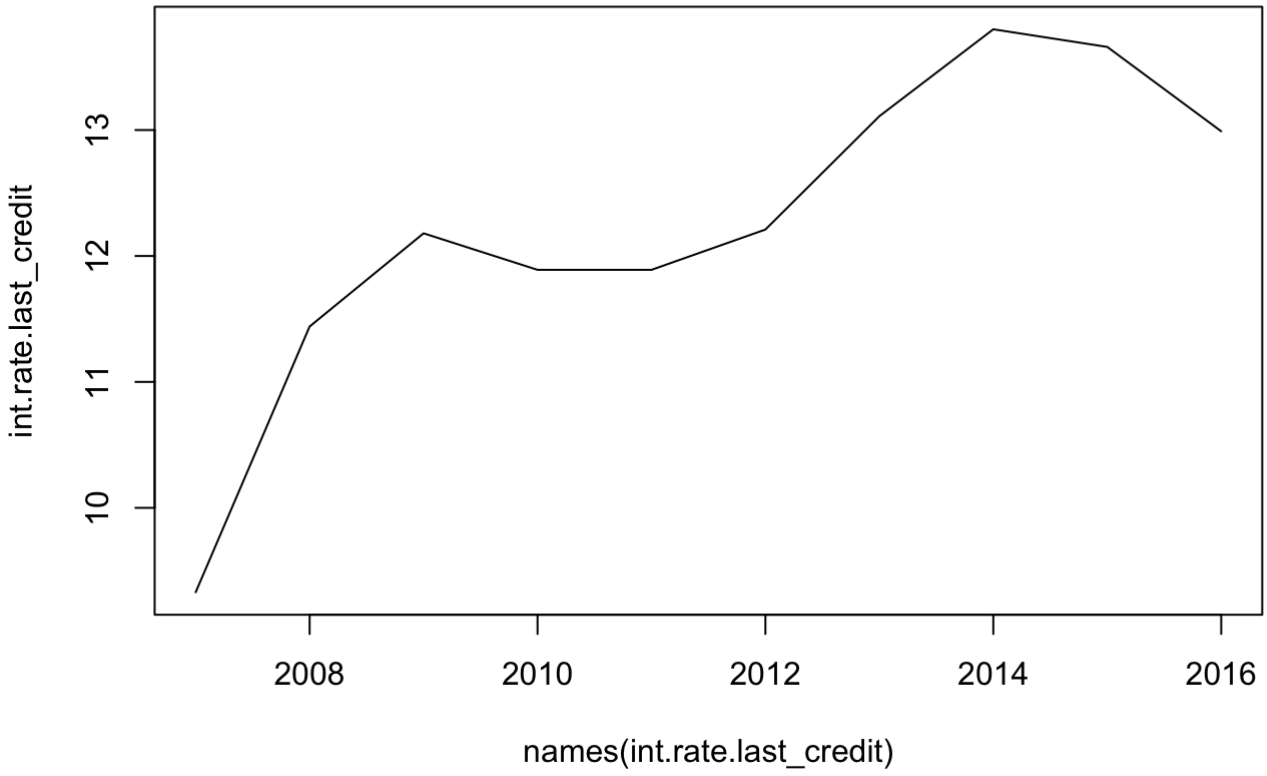
```
earliest_cr_line_date <-as.Date(as.yearmon(loan$earliest_cr_line,
                                           '%b-%Y'))
loan$earliest_cr_line_year<- format(earliest_cr_line_date, '%Y')
int.rate.earliest <- by(loan, loan$earliest_cr_line_year,
                     function(x){return(median(x$int_rate))})
plot(names(int.rate.earliest), int.rate.earliest, type = 'l')
```




```
#21"last_credit_pull_d" [reserve]
head(loan[,"last_credit_pull_d"])
```

```
## [1] "Jan-2016" "Sep-2013" "Jan-2016" "Jan-2015" "Jan-2016" "Sep-2015"
```

```
last_credit_pull_d.date <-as.Date(as.yearmon(loan$last_credit_pull_d,
                                             '%b-%Y'))
loan$last_credit_pull_year<- format(last_credit_pull_d.date, '%Y')
int.rate.last_credit <- by(loan, loan$last_credit_pull_year,
                           function(x){return(median(x$int_rate))})
plot(names(int.rate.last_credit), int.rate.last_credit, type = 'l')
```




```
#22"application_type"
table(loan[,"application_type"]) #not considered
```

```
##
## INDIVIDUAL      JOINT
##     886868        511
```

```
#23"verification_status_joint"
table(loan[,"verification_status_joint"]) #not consider too much missing
```

```
##
##                    Not Verified Source Verified        Verified
##          886868             283              61             167
```

reserved being considered are: “term”,“sub_grade”, “home_ownership”, “verification_status”,“issue_d”, “purpose”, “earliest_cr_line”, “last_credit_pull_d”

5 catergorical variables that influence interest rate most are:

“term”,“sub_grade”, “home_ownership”,“issue_year”,“purpose”