Lending Club 2 Dataset

Nettoyage des données

```
##
## Attaching package: 'lubridate'
## The following object is masked from 'package:base':
##
##     date
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:lubridate':
##
##     intersect, setdiff, union
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Load du fichier

```
dirpath <- "C:/Users/sebde/OneDrive/Documents/cepe-2018-kickclub/Data"
filename <- "loan.csv"
my_file <- paste(dirpath, filename, sep = "/")
```

On commence par charger les donnees originales

```
loan5 <- read.csv(my_file, header = TRUE, sep = ",", dec = ".")
```

On nettoye les dates

```
## Conversion des dates format 'factor' en dates format 'Date'
dt <- c("issue_d","earliest_cr_line")

period_factor <- function(x){

  ## convertir en char
  x.char <- as.character(x)

  ## ajouter un jour pour que la chaîne de caractère soit ensuite reconnu comme date
  x.char <- paste("01-", x.char, sep = "")
```

```
  ## convertir en date
  Sys.setlocale("LC_TIME", "English_United States")
  x.date <- as.Date(x.char, format = c("%d-%b-%Y"))

  return(x.date)

}

for (col in dt) {
  loan5[,col] <- period_factor(loan5[,col])
}
```

Creation de nouvelles variables

- "length_cr_line": la diférence entre "issue_d" et "earliest_cr_line", soit l'historique de crédit au moment de l'émission du pret

```
loan5 <- loan5 %>% mutate(length_cr_line = as.numeric(issue_d - earliest_cr_line))
```

```
## Warning: package 'bindrcpp' was built under R version 3.4.4
```

```
summary(loan5$length_cr_line)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     184    4110    5419    5970    7336   25933      29
```

- "issue_y" l'ann?e plut?t que le mois ("issue_d") d'?mission du pret

```
loan5 <- loan5 %>% mutate(issue_y = factor(year(issue_d)))
summary(loan5$issue_y)
```

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

On garde uniquement les modalités qui nous interessent

- Fully Paid: le crédit est arrivé à terme
- Charged Off: l'emprunteur est en cessation de paiement

Les autres modalités sont enlevées (prêt en cours)

```
## ceux qui ont payé jusqu'au bout
paid <- c("Fully Paid")

## ceux qui sont en "faillite"
not_paid <- c("Charged Off")

loan5 <- loan5[loan5$loan_status %in% c(paid, not_paid),]
loan5$loan_status <- as.factor(as.character(loan5$loan_status))
```

Des variables sont fortement corrélées

Une trop forte corrélation entre les variables fait parfois planter certains modèles

```
cor(loan5[,c("loan_amnt", "funded_amnt", "funded_amnt_inv")])
```

```
##                 loan_amnt funded_amnt funded_amnt_inv
## loan_amnt       1.0000000   0.9976053       0.9917406
## funded_amnt     0.9976053   1.0000000       0.9946117
## funded_amnt_inv 0.9917406   0.9946117       1.0000000
```

Liste des variables selectionnées

Après analyse fonctionnelle des variables:

- taux de NA trop élevé
- variables connues a posteriori

On décide de garder les variables suivantes:

```
keep <- c("loan_status", "funded_amnt", "term", "int_rate", "installment", "grade", "sub_grade", "emp_l
keep
```

```
##  [1] "loan_status"         "funded_amnt"         "term"
##  [4] "int_rate"            "installment"         "grade"
##  [7] "sub_grade"           "emp_length"          "home_ownership"
## [10] "annual_inc"          "verification_status" "purpose"
## [13] "addr_state"          "dti"                 "delinq_2yrs"
## [16] "inq_last_6mths"      "open_acc"            "pub_rec"
## [19] "revol_bal"           "revol_util"          "total_acc"
## [22] "initial_list_status" "acc_now_delinq"      "length_cr_line"
## [25] "issue_y"
```

Selection des variables dans le dataset

```
loan5 <- loan5[,keep]
```

Traitement supplementaire des donnees

Apercu du jeu de donnees selectionnees

```
summary(loan5)
```

```
##       loan_status      funded_amnt            term           int_rate
##  Charged Off: 45248   Min.   :  500    36 months:196658   Min.   : 5.32
##  Fully Paid :207723   1st Qu.: 7200    60 months: 56313   1st Qu.:10.74
##                       Median :12000                       Median :13.53
##                       Mean   :13522                       Mean   :13.76
##                       3rd Qu.:18075                       3rd Qu.:16.55
##                       Max.   :35000                       Max.   :28.99
##
##   installment      grade       sub_grade          emp_length
##  Min.   :  15.69   A:42296   B3     : 18068   10+ years:76881
##  1st Qu.: 239.55   B:76065   B4     : 16933   2 years  :23561
##  Median : 365.23   C:65320   C1     : 14959   < 1 year :20886
##  Mean   : 418.11   D:40506   B2     : 14628   3 years  :20380
##  3rd Qu.: 547.55   E:19186   C2     : 14341   5 years  :18059
```

```
##  Max.   :1424.57   F: 7660   B5     : 14340   1 year   :16856
##                    G: 1938   (Other):159702   (Other)  :76348
##   home_ownership     annual_inc           verification_status
##  ANY     :     1   Min.   :   3000   Not Verified   :86064
##  MORTGAGE:124844   1st Qu.:  45000   Source Verified:74011
##  NONE    :    43   Median :  62000   Verified       :92896
##  OTHER   :   141   Mean   :  72538
##  OWN     : 21985   3rd Qu.:  87000
##  RENT    :105957   Max.   :8706582
##
##                purpose         addr_state          dti
##  debt_consolidation:148363   CA     : 43110   Min.   : 0.00
##  credit_card       : 50076   NY     : 21338   1st Qu.:10.75
##  home_improvement  : 14929   TX     : 19343   Median :16.20
##  other             : 14277   FL     : 17545   Mean   :16.54
##  major_purchase    :  6265   NJ     :  9601   3rd Qu.:21.99
##  small_business    :  4746   IL     :  9253   Max.   :57.14
##  (Other)           : 14315   (Other):132781
##   delinq_2yrs      inq_last_6mths      open_acc        pub_rec
##  Min.   : 0.0000   Min.   :0.0000   Min.   : 0.00   Min.   : 0.0000
##  1st Qu.: 0.0000   1st Qu.:0.0000   1st Qu.: 7.00   1st Qu.: 0.0000
##  Median : 0.0000   Median :1.0000   Median :10.00   Median : 0.0000
##  Mean   : 0.2499   Mean   :0.8525   Mean   :10.94   Mean   : 0.1434
##  3rd Qu.: 0.0000   3rd Qu.:1.0000   3rd Qu.:14.00   3rd Qu.: 0.0000
##  Max.   :29.0000   Max.   :8.0000   Max.   :76.00   Max.   :15.0000
##
##    revol_bal         revol_util       total_acc      initial_list_status
##  Min.   :      0   Min.   :  0.00   Min.   :  2.00   f:182079
##  1st Qu.:   5862   1st Qu.: 36.30   1st Qu.: 16.00   w: 70892
##  Median :  10937   Median : 55.80   Median : 23.00
##  Mean   :  15168   Mean   : 54.31   Mean   : 25.04
##  3rd Qu.:  19067   3rd Qu.: 73.90   3rd Qu.: 32.00
##  Max.   :1746716   Max.   :892.30   Max.   :150.00
##                    NA's   :199
##  acc_now_delinq     length_cr_line     issue_y
##  Min.   :0.000000   Min.   : 1095   2013   :71232
##  1st Qu.:0.000000   1st Qu.: 3865   2014   :68694
##  Median :0.000000   Median : 5084   2012   :49563
##  Mean   :0.003115   Mean   : 5570   2015   :25757
##  3rd Qu.:0.000000   3rd Qu.: 6820   2011   :19675
##  Max.   :5.000000   Max.   :24138   2010   :11521
##                                     (Other): 6529
```

- Nettoyage des derniers NAs

```
loan5 <- filter(loan5, !is.na(loan5[,"revol_util"]))
```

- Conversion de la variable 'home_owner_ship' a trois modalites seulement

```
loan5 <- loan5 %>% filter(home_ownership == "MORTGAGE" | home_ownership == "RENT" | home_ownership == "
loan5$home_ownership <- factor(loan5$home_ownership) ## delete the unused levels with no more observati
```

- Relevel des loan_status: variable à expliquer

```
levels(loan5$loan_status) <- c("CO", "FP")
```

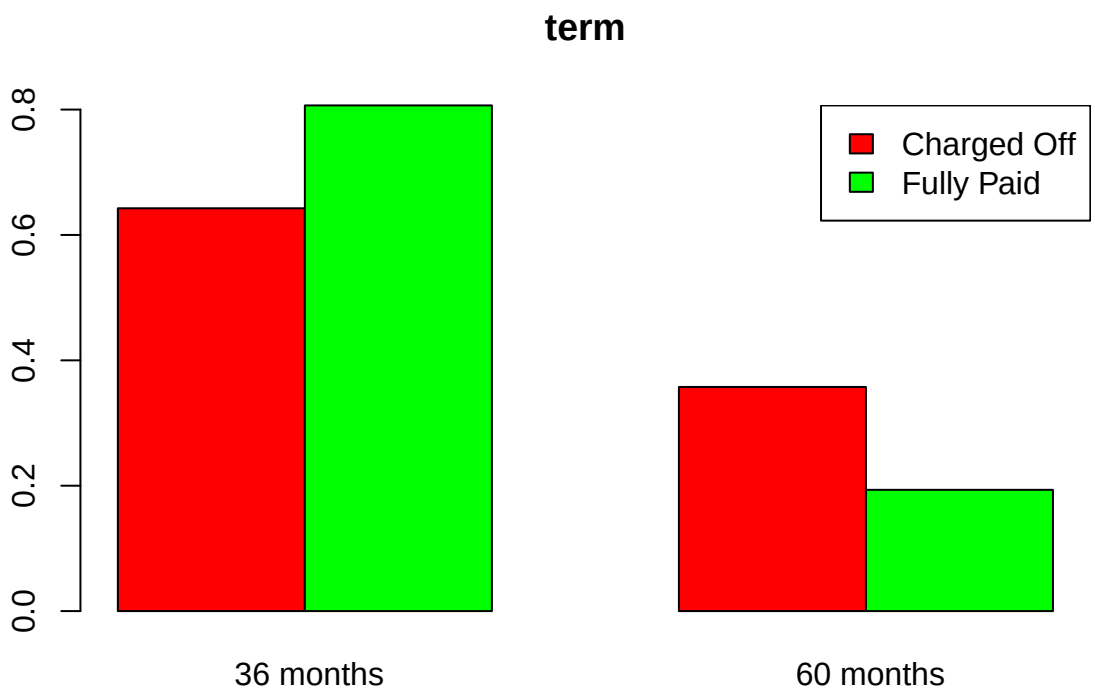
Sauvegarder le jeu de donnees pour la modelisation

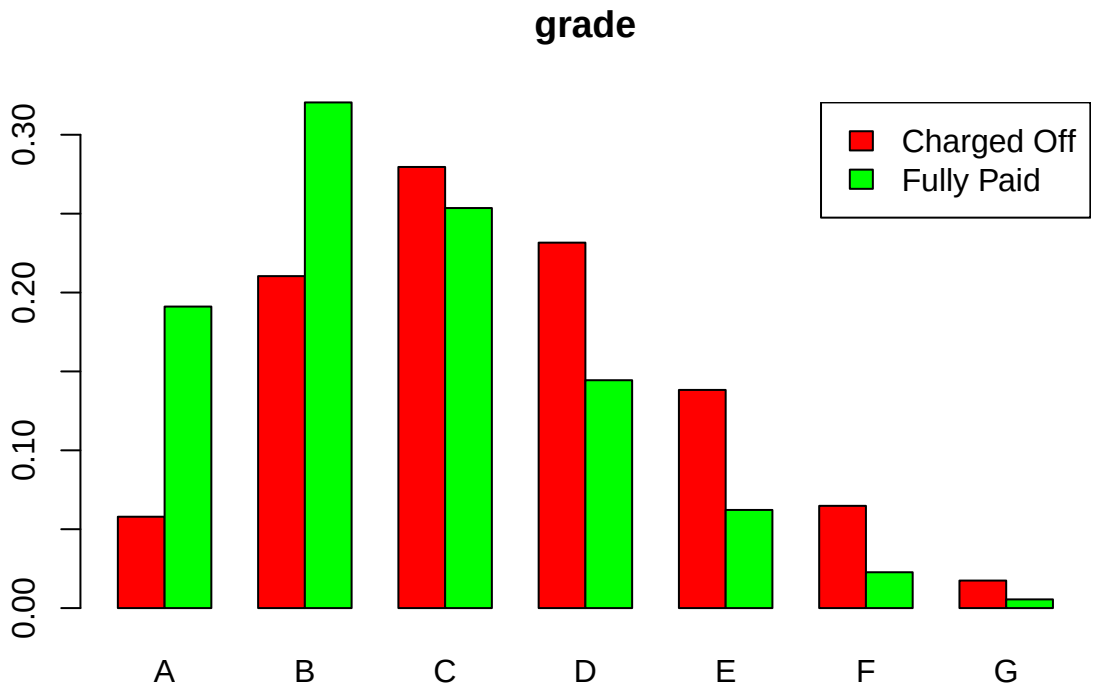
```
clean_file_name <-  "loan5.RDS"
saveRDS(object = loan5, file = paste(dirpath, clean_file_name, sep = "/"))
```

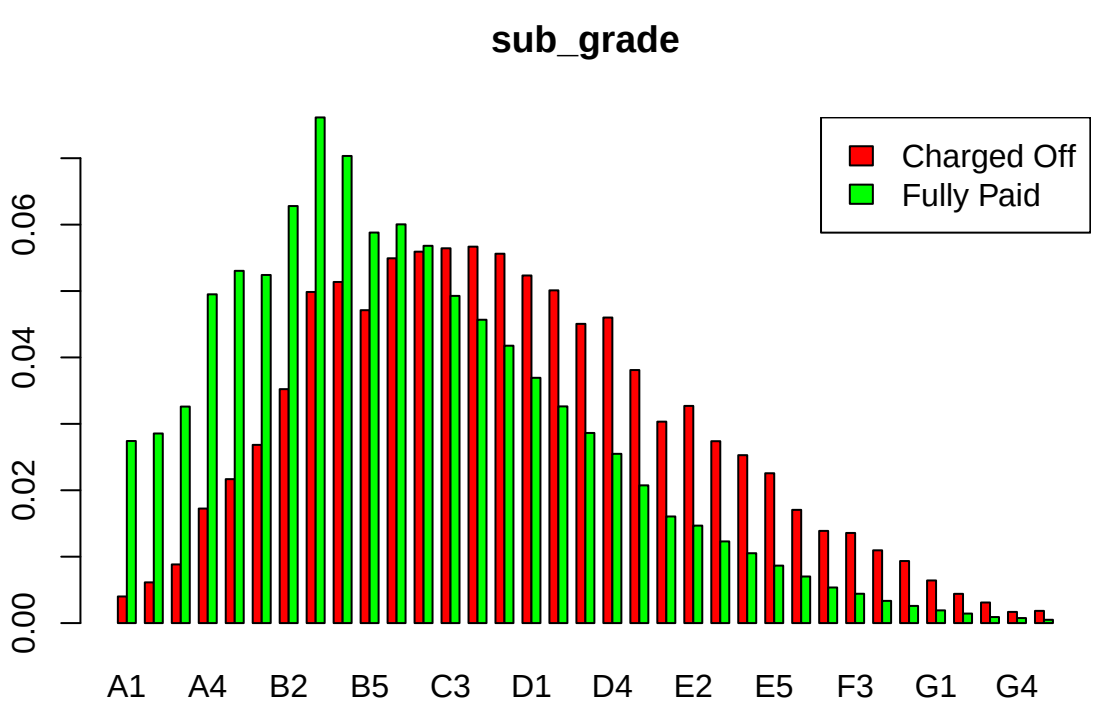
Visualisation des variables conservées

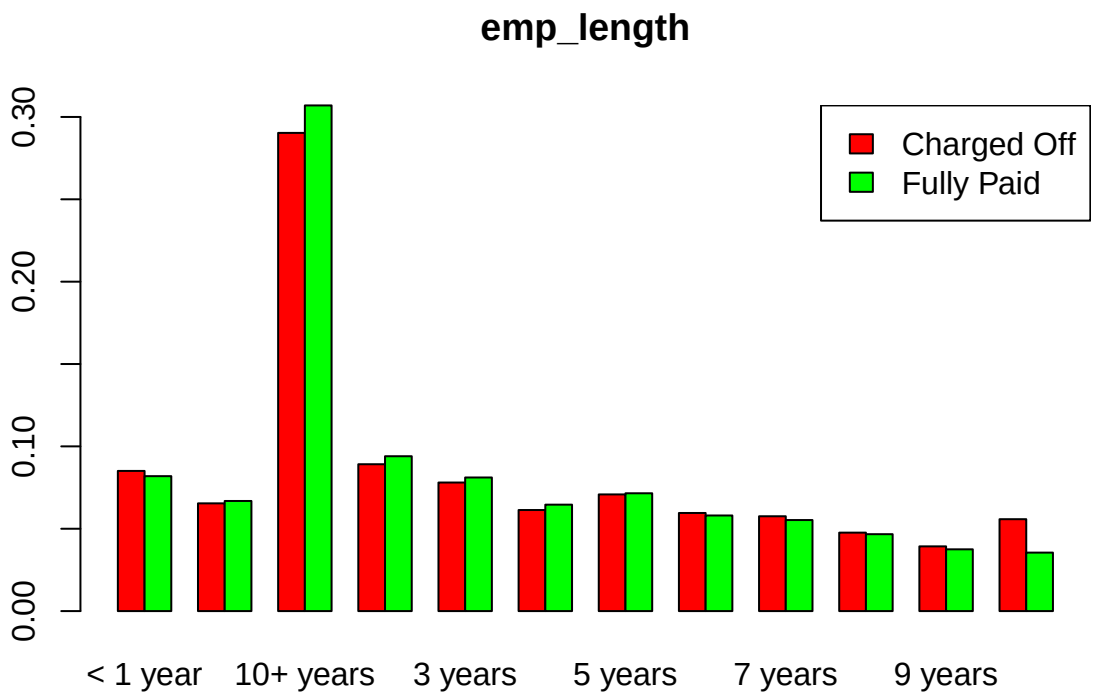
- Variables 'factor'

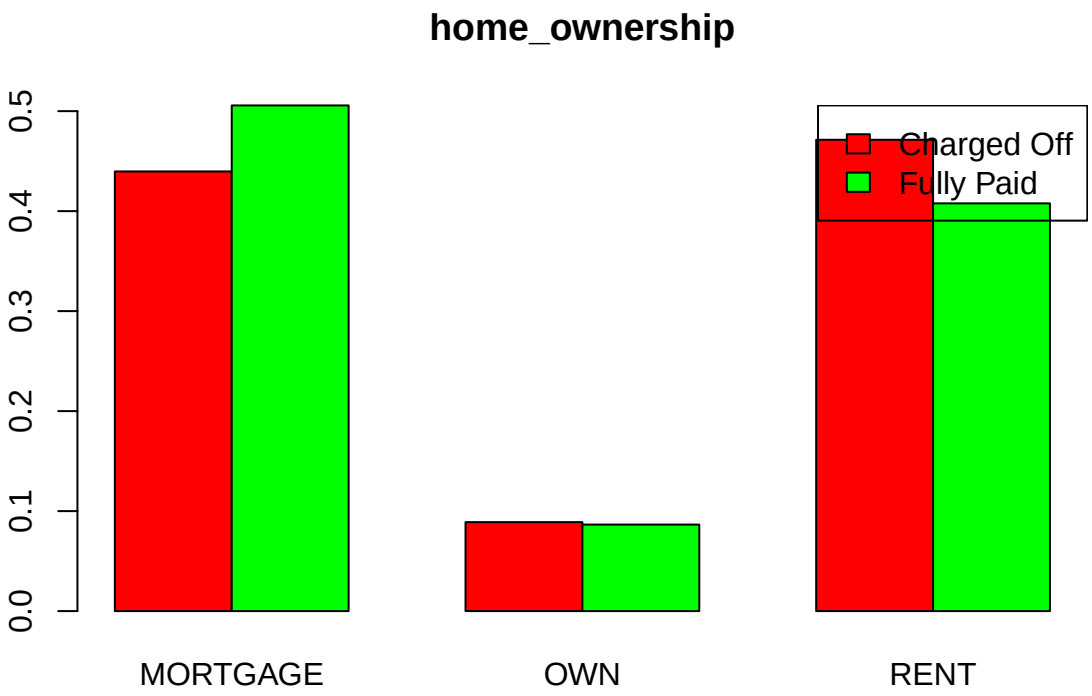
```
## on conserve le nombre dans chaque niveau pour pouvoir ensuite plutôt bosser sur des pourcentages
nb_not_paid <- nrow(loan5[loan5$loan_status == "CO",])
nb_paid <- nrow(loan5[loan5$loan_status == "FP",])

## affichage des factor pour voir
loan5.class <- sapply(loan5, class)
mod.alt <- names(which(loan5.class == "factor"))

for (i in mod.alt) {
  if (i != "loan_status") {
    t <- table(loan5[,c("loan_status", i)])
    t["CO",] <- t["CO",]/nb_not_paid
    t["FP",] <- t["FP",]/nb_paid

    barplot(t, beside = TRUE, col = c("red","green"), main = i)
    legend("topright", legend = c("Charged Off", "Fully Paid"), fill = c("red", "green"))
  }
}
```

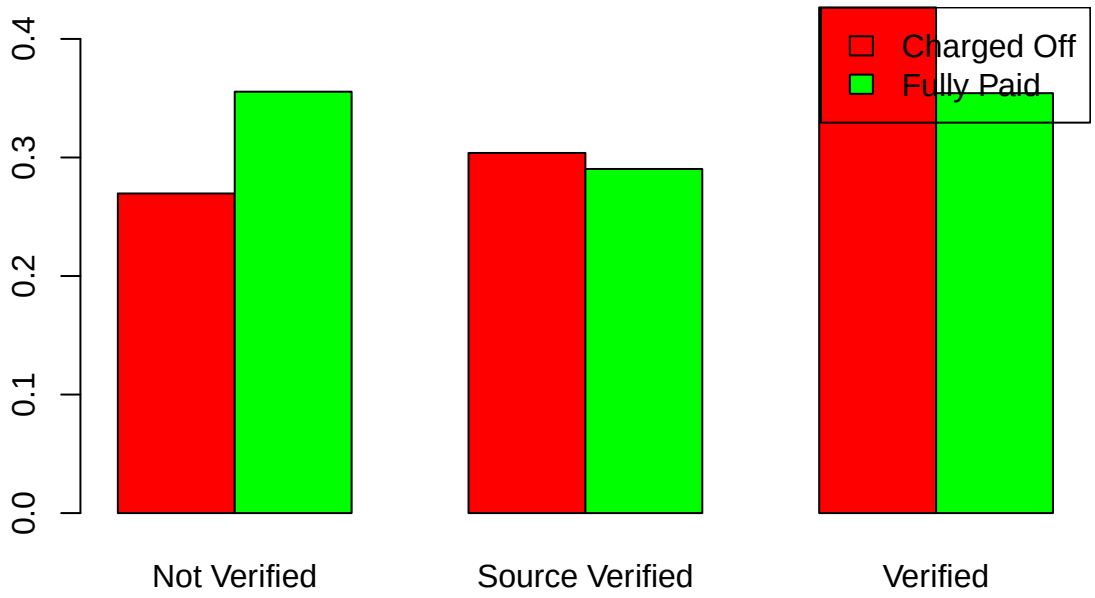

term
0.0
0.2
0.4
0.6
0.8
Charged Off
Fully Paid
36 months
60 months

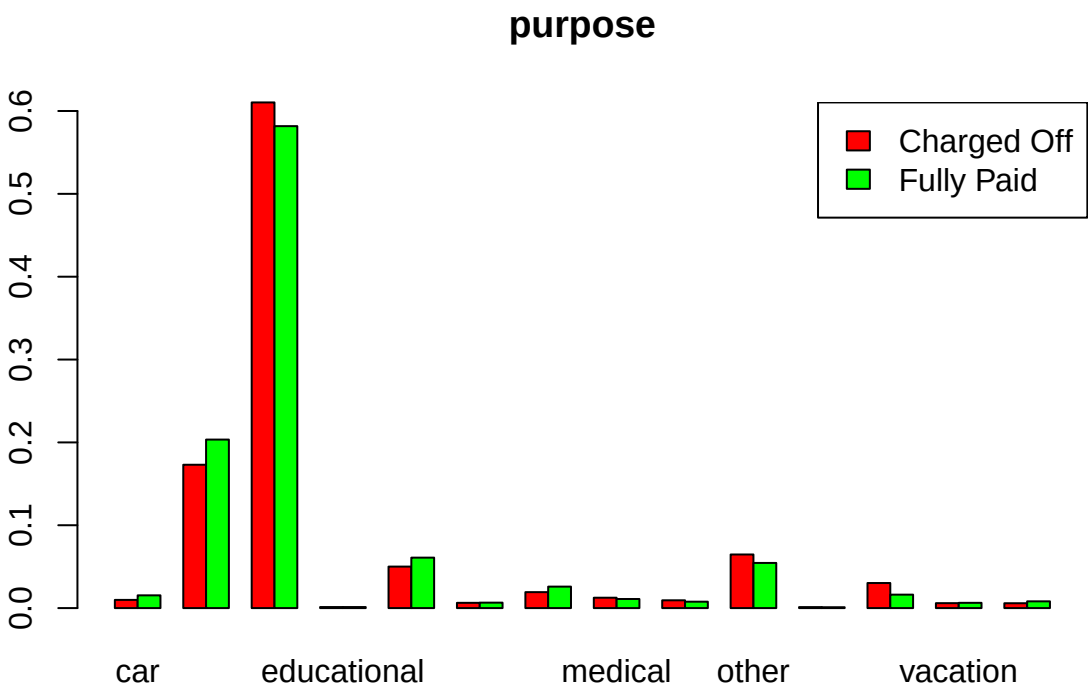

grade
Charged Off
Fully Paid
0.00
0.10
0.20
0.30
A
B
C
D
E
F
G

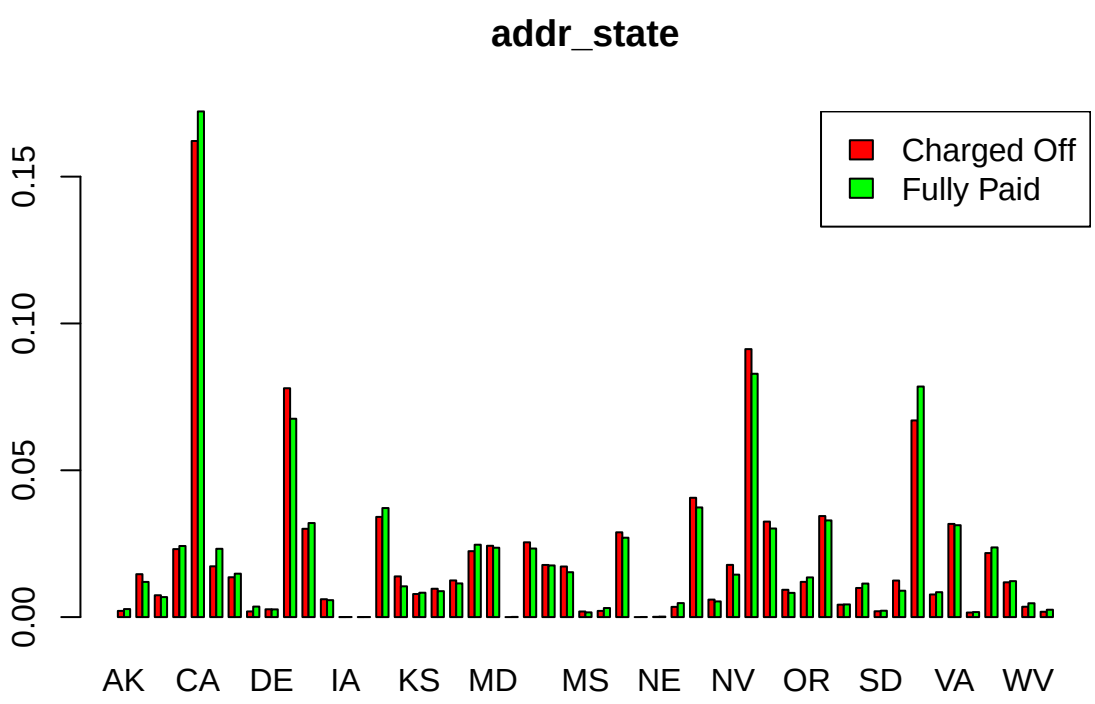

sub_grade
Charged Off
Fully Paid
0.00
0.02
0.04
0.06
A1
A4
B2
B5
C3
D1
D4
E2
E5
F3
G1
G4

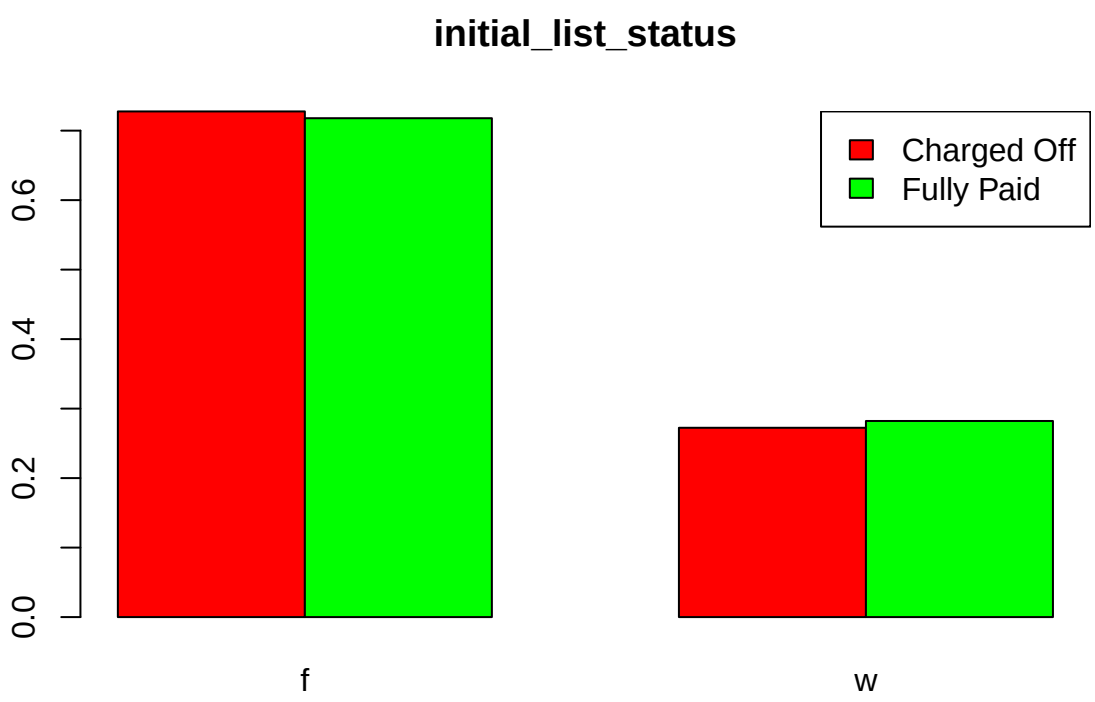

emp_length
0.00
0.10
0.20
0.30
Charged Off
Fully Paid
< 1 year
10+ years
3 years
5 years
7 years
9 years

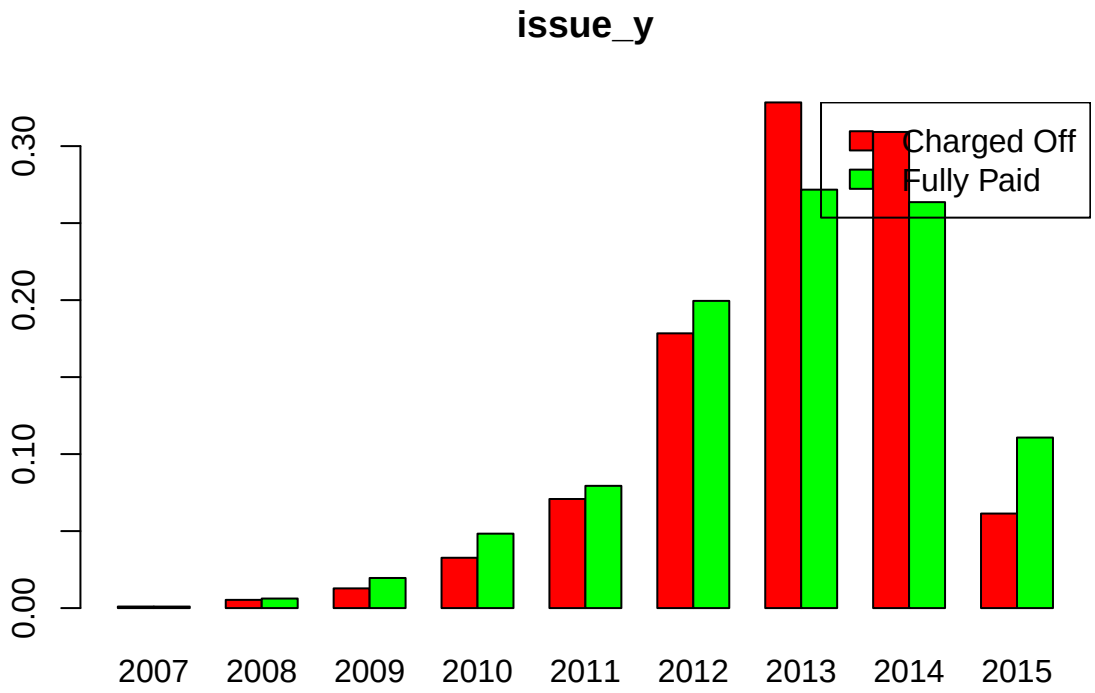

home_ownership
0.0
0.1
0.2
0.3
0.4
0.5
Charged Off
Fully Paid
MORTGAGE
OWN
RENT


verification_status
0.0
0.1
0.2
0.3
0.4
Charged Off
Fully Paid
Not Verified
Source Verified
Verified


purpose
Charged Off
Fully Paid
0.0
0.1
0.2
0.3
0.4
0.5
0.6
car
educational
medical
other
vacation

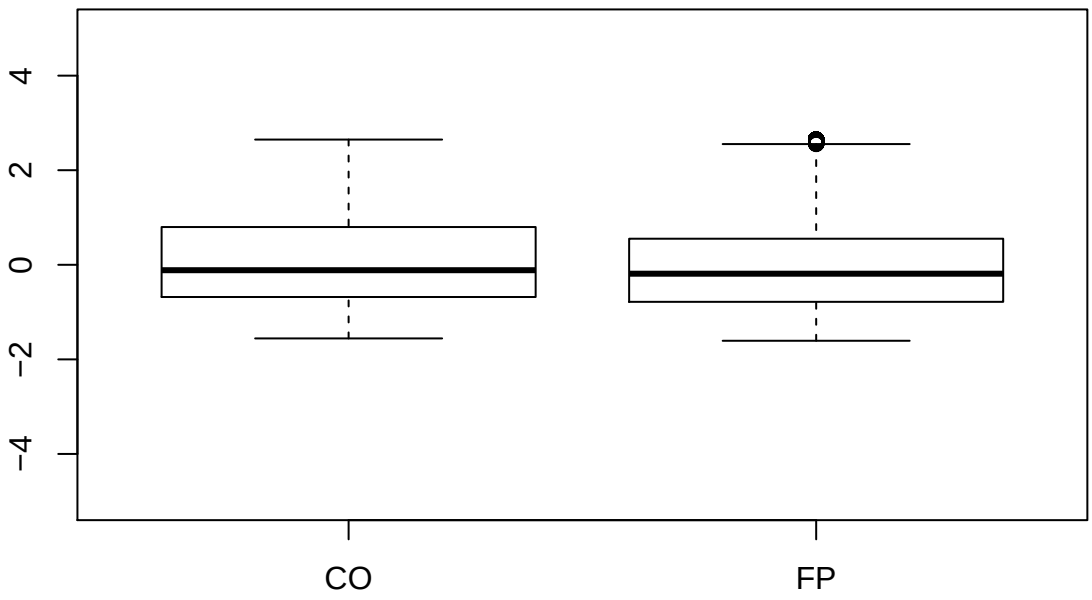

addr_state
Charged Off
Fully Paid
0.00
0.05
0.10
0.15
AK
CA
DE
IA
KS
MD
MS
NE
NV
OR
SD
VA
WV

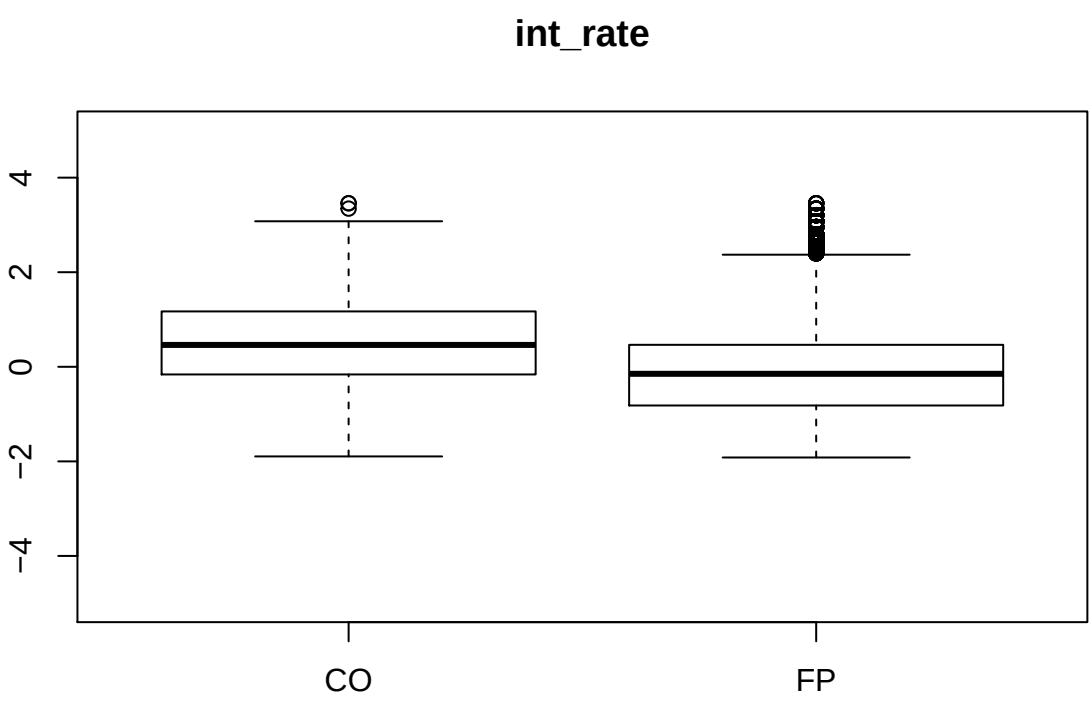

initial_list_status
Charged Off
Fully Paid
0.0
0.2
0.4
0.6
f
w

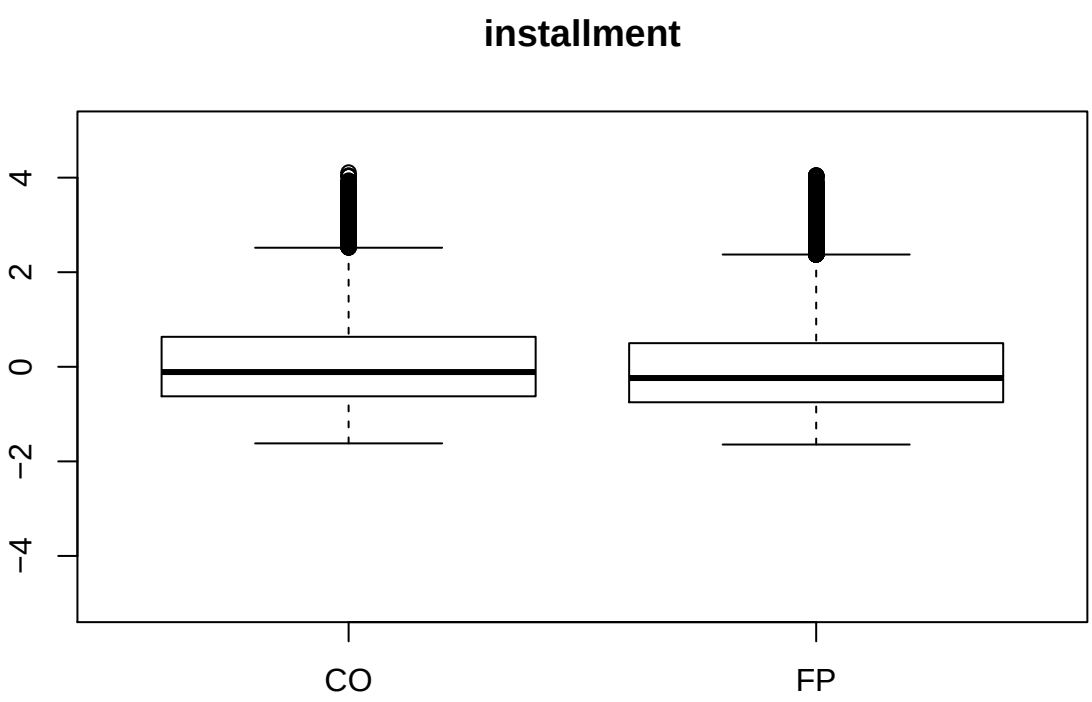


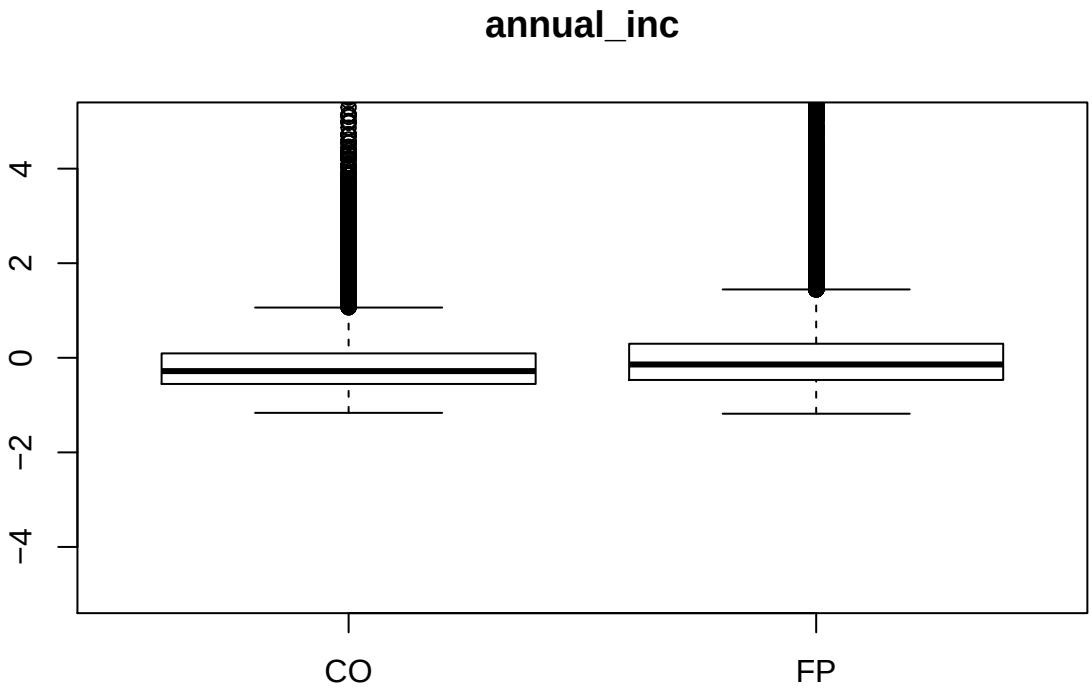
Variables 'numeric'

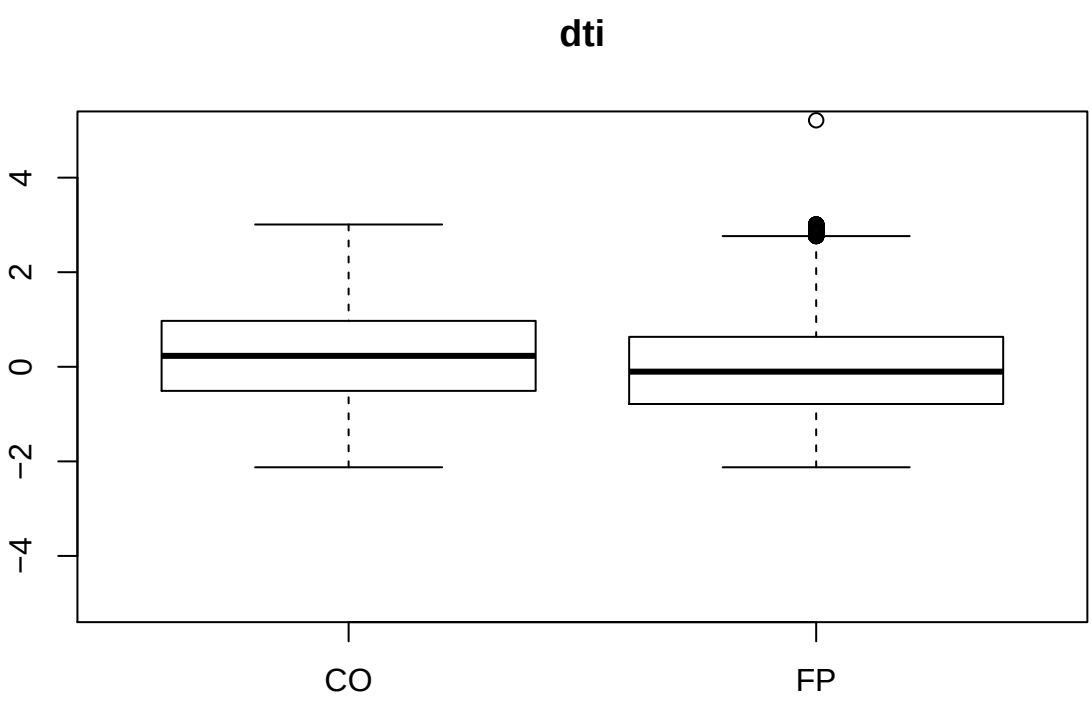
```
## affichage des num pour voir
num.alt <- which(loan5.class %in% c("integer","numeric"))

for (i in num.alt) {
  if (i != "loan_status") {
    boxplot(scale(loan5[,i])~loan5$loan_status, main = colnames(loan5[i]), ylim = c(-5,5))
  }
}
```

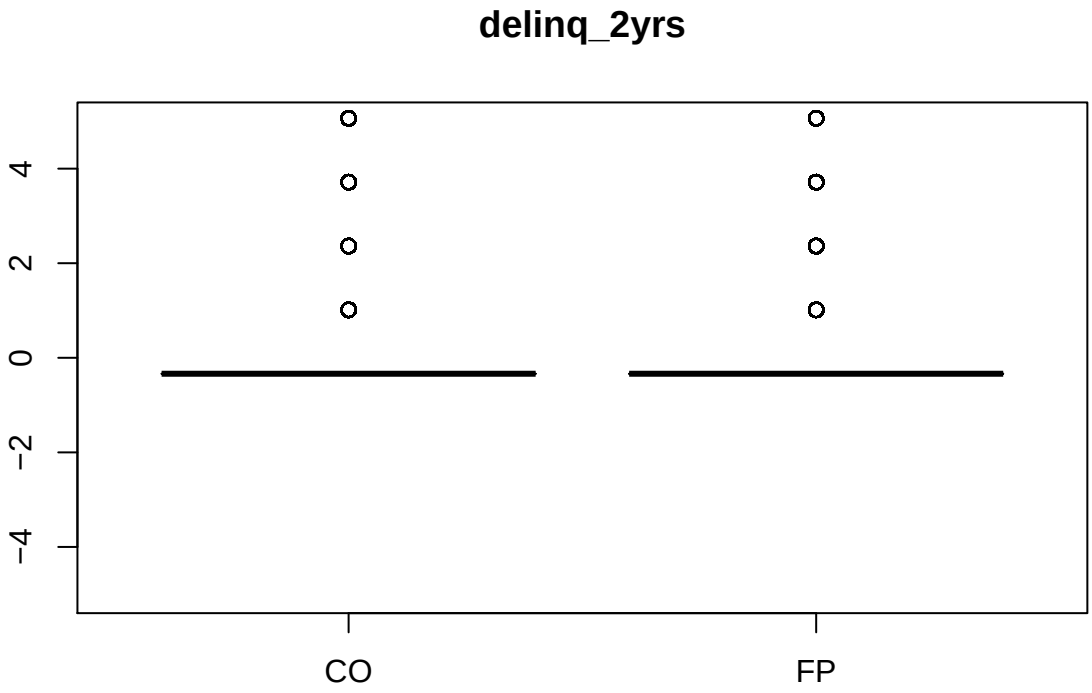

funded_amnt
4
2
0
−2
−4
CO
FP

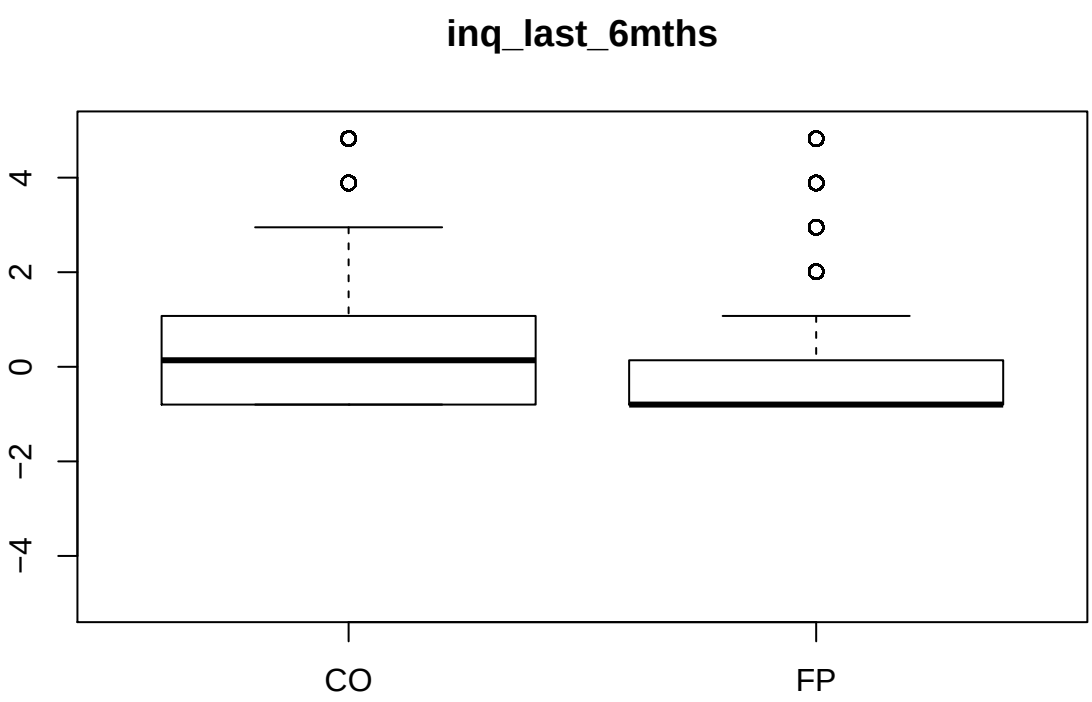

int_rate
4
2
0
−2
−4
CO
FP

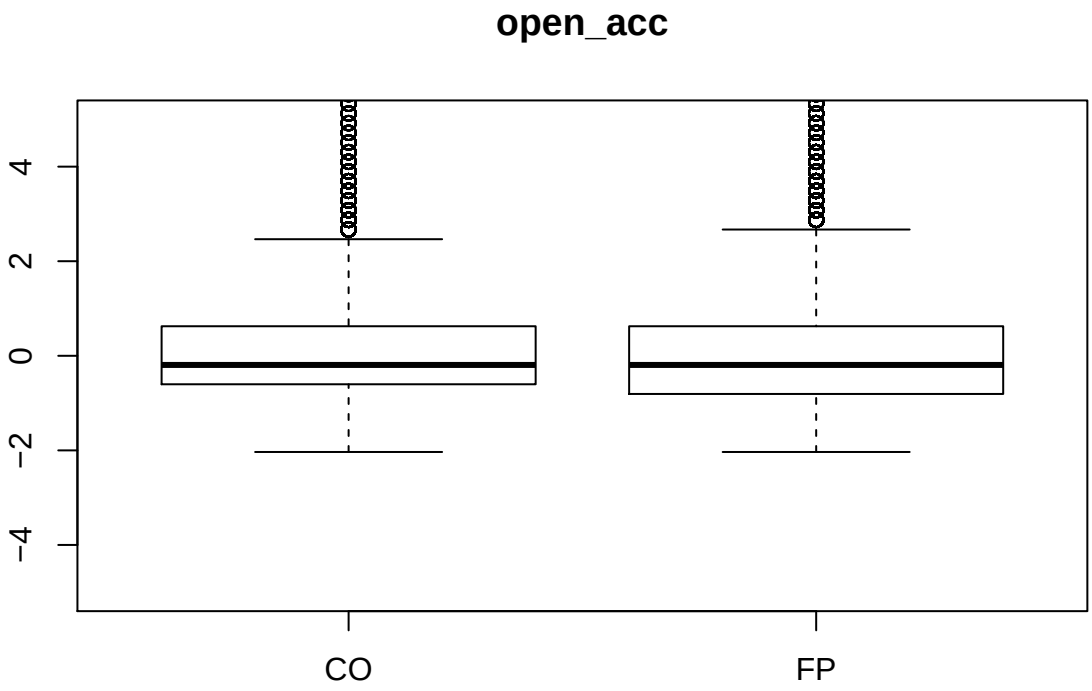

installment
4
2
0
−2
−4
CO
FP

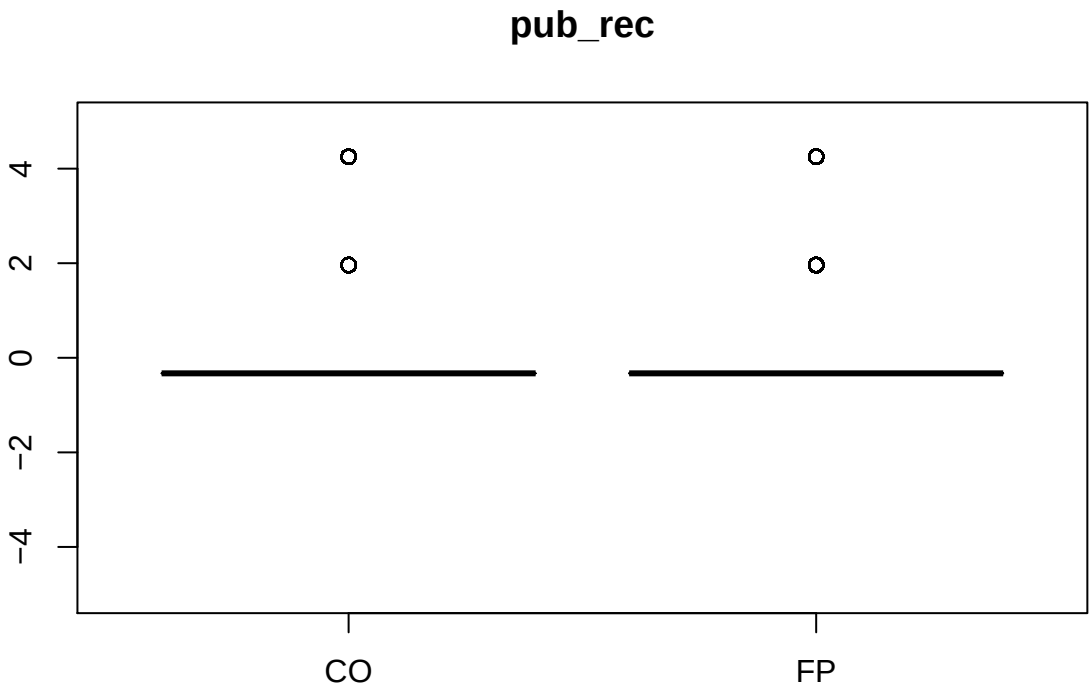

annual_inc
4
2
0
−2
−4
CO
FP

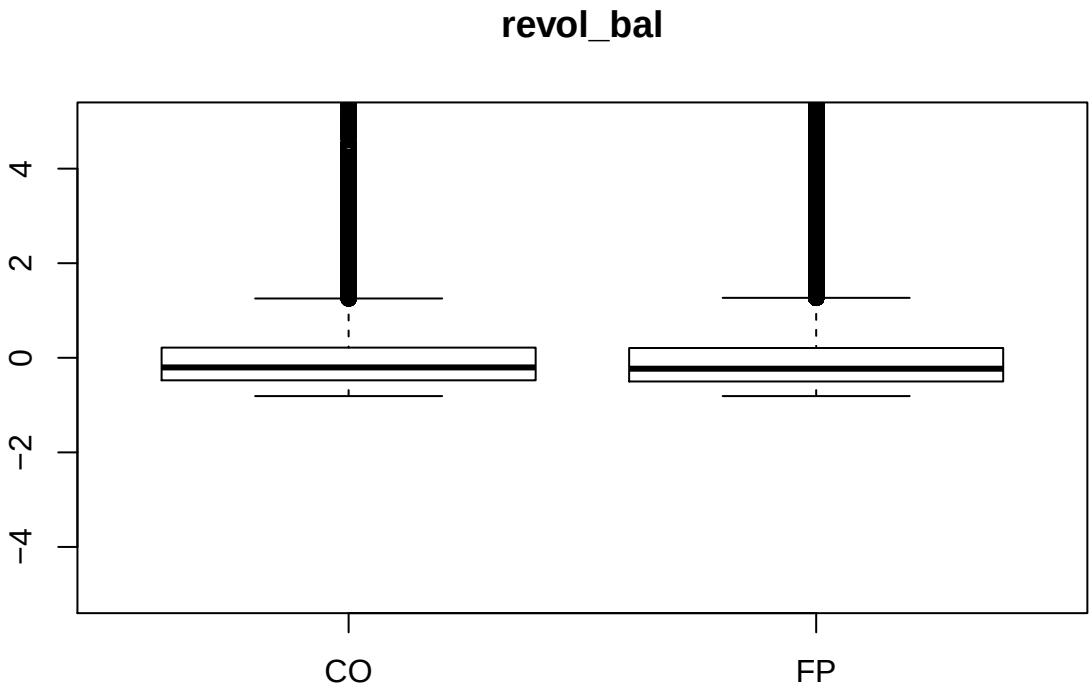

dti
4
2
0
−2
−4
CO
FP

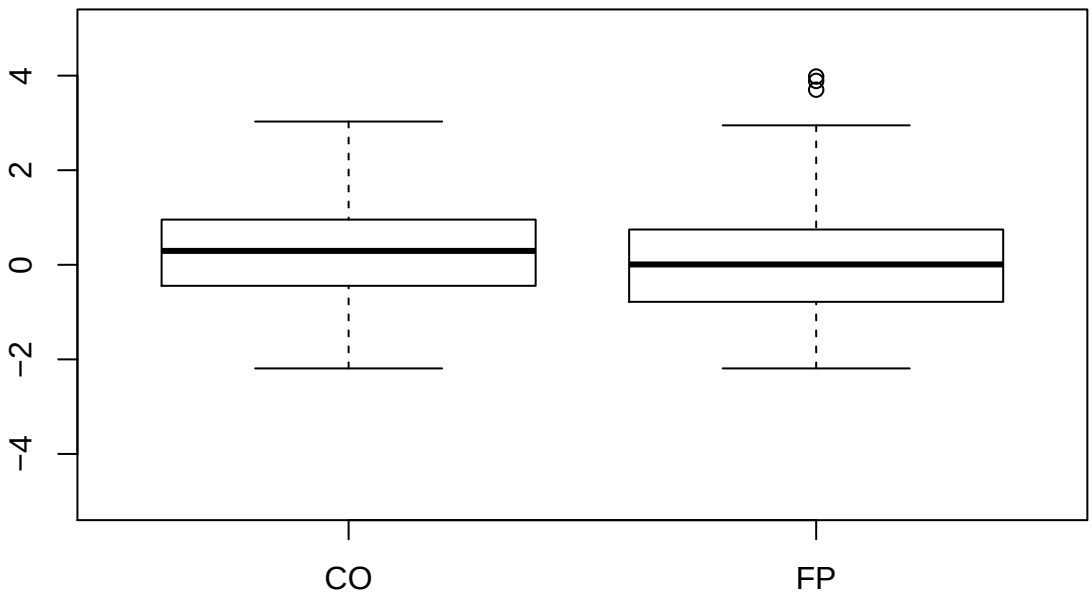

delinq_2yrs
4
2
0
−2
−4
CO
FP

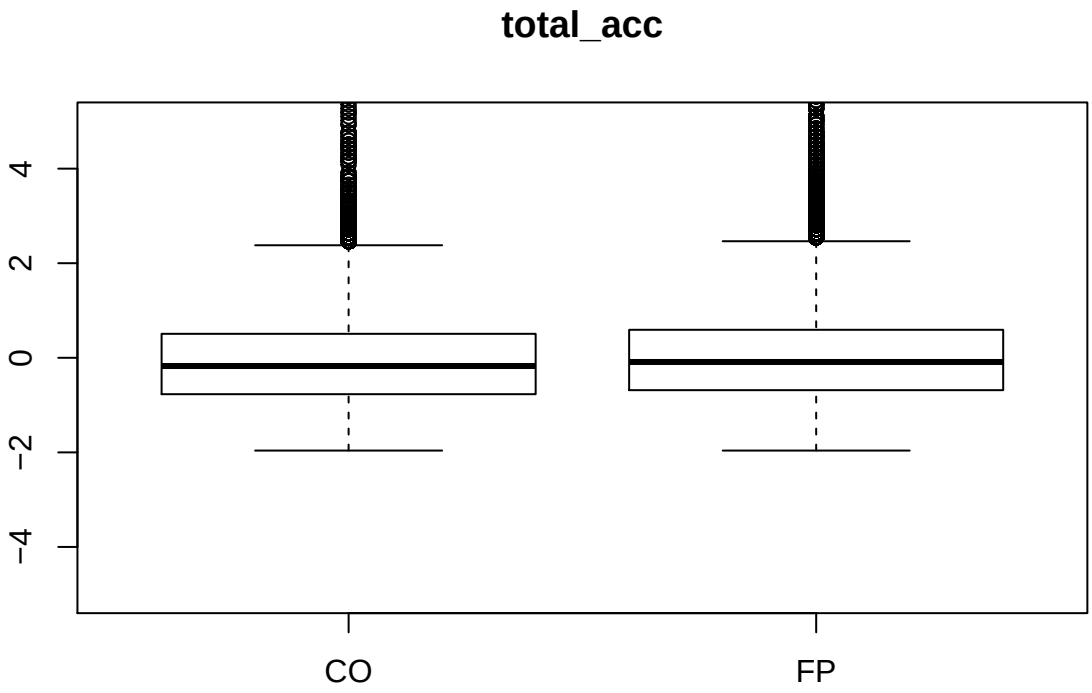

inq_last_6mths
4
2
0
−2
−4
CO
FP

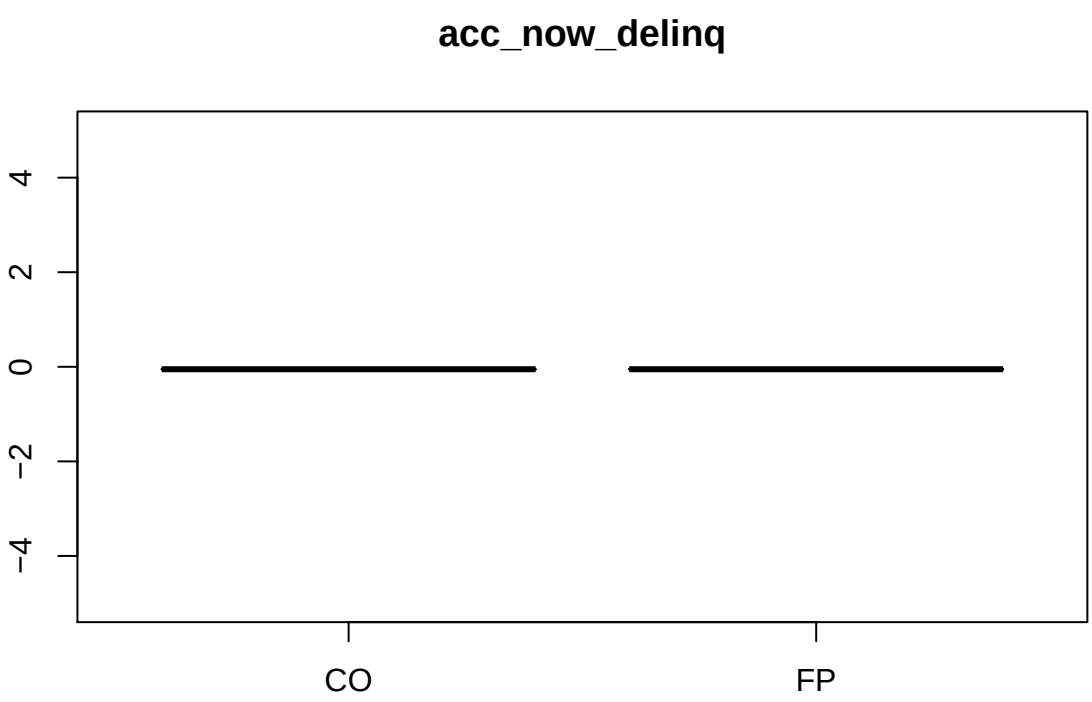

open_acc
4
2
0
−2
−4
CO
FP

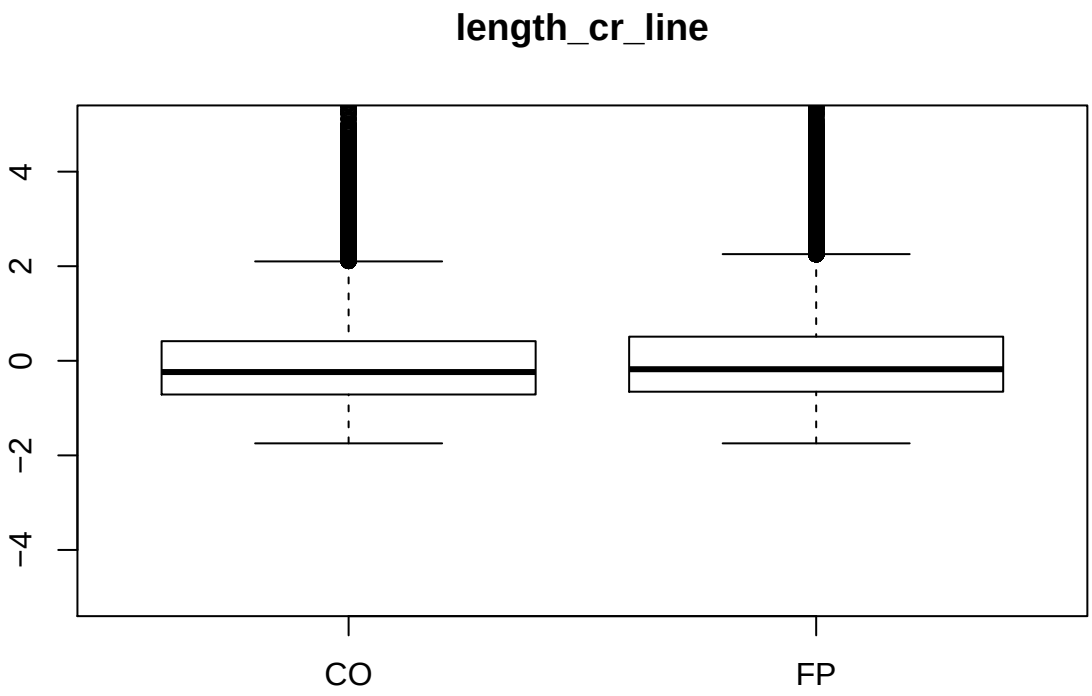

pub_rec
4
2
0
−2
−4
CO
FP


revol_bal
4
2
0
−2
−4
CO
FP


revol_util
4
2
0
−2
−4
CO
FP


total_acc
4
2
0
−2
−4
CO
FP


acc_now_delinq
4
2
0
−2
−4
CO
FP


length_cr_line
4
2
0
−2
−4
CO
FP