dataPreprocessing

```
library(caret)
```

```
## Loading required package: lattice
## Loading required package: ggplot2
```

```
library(rpart)
library(rattle)
```

```
## Rattle: A free graphical interface for data mining with R.
## Version 3.4.1 Copyright (c) 2006-2014 Togaware Pty Ltd.
## Type 'rattle()' to shake, rattle, and roll your data.
```

```
library(corrplot)
library(car)
data = read.csv("../documents/JHU-algo/Homework/OnDeck Analytics
Asssignment.csv",header=TRUE)
```

```
data$diff = data$days_delinquent_new - data$days_delinquent_old #the larger
the worse
data$target[data$diff<=0]=0 #remain the same or geting better
data$target[data$diff>0]=1  #geting worse
data$lender1[data$lender_payoff == 0] <- 0
data$lender1[data$lender_payoff > 0] <- 1

#noticed missing values
summary(data)
```

```
##    as_of_date  days_delinquent_old days_delinquent_new
##  11/1/12:477   Min.   :  1.00      Min.   :  0.00
##                1st Qu.:  4.00      1st Qu.:  5.00
##                Median : 14.00      Median : 22.00
##                Mean   : 26.66      Mean   : 32.55
##                3rd Qu.: 38.00      3rd Qu.: 51.00
##                Max.   :180.00      Max.   :180.00
##
##  new_outstanding_principal_balance initial_loan_amount      fico
##  Min.   :    -0.15                 Min.   :  6000      Min.   :501.0
##  1st Qu.:  7771.34                 1st Qu.: 15000      1st Qu.:591.0
##  Median : 15817.58                 Median : 25000      Median :641.0
##  Mean   : 21406.54                 Mean   : 31448      Mean   :637.3
##  3rd Qu.: 29222.66                 3rd Qu.: 40000      3rd Qu.:683.0
##  Max.   :125000.00                 Max.   :135000      Max.   :806.0
##                                                        NA's   :1
##                          sales_channel__c                  type
##  Direct                           : 74     Loan - New Customer:352
##  FAP: Managed Application Program:370     Loan - Renewal     :125
##  Promontory                       :  1
```

```
##  Referral                        : 32
##
##
##
##             current_collection_method      term        lender_payoff
##  ACH Pull                 :423         Min.   : 3.000   Min.   :    0
##  Split Funding            : 35         1st Qu.: 6.000   1st Qu.:    0
##  Transfer Account Vendors: 19         Median : 6.000   Median :    0
##                                       Mean   : 8.205   Mean   : 1375
##                                       3rd Qu.:12.000   3rd Qu.:    0
##                                       Max.   :18.000   Max.   :38691
##
##  average_bank_balance__c last_cleared_payment_date      diff
##  Min.   :   357.1        11/1/12 :211              Min.   :-180.000
##  1st Qu.:  2967.1        10/31/12: 38              1st Qu.:   0.000
##  Median :  5717.4        10/30/12: 17              Median :   9.000
##  Mean   : 11632.2        10/29/12: 14              Mean   :   5.893
##  3rd Qu.: 11103.0        10/26/12: 11              3rd Qu.:  19.000
##  Max.   :340080.2        10/18/12:  8              Max.   :  19.000
##  NA's   :1               (Other) :178
##      target          lender1
##  Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000
##  Median :1.0000   Median :0.0000
##  Mean   :0.6583   Mean   :0.1195
##  3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000
##
```

Inpute missing values

```
# FICO score is missing on row 7, we replace it with the median value.
data$fico[7] <- median(data$fico, na.rm=TRUE)

# one missing record on row 137 for the average_bank_balance, we can replace
it with mean, median, or just impute it based on anova method.
predicted_balance <- rpart(average_bank_balance__c ~
new_outstanding_principal_balance + initial_loan_amount +
fico,data=data[!is.na(data$average_bank_balance__c),], method="anova")

data$average_bank_balance__c[is.na(data$average_bank_balance__c)] <-
predict(predicted_balance, data[is.na(data$average_bank_balance__c),])
inTrain = createDataPartition(y=data$target, p=0.75, list=FALSE)
training = data[inTrain,]
testing = data[-inTrain,]
dim(training);dim(testing)

## [1] 358  16
```

```
## [1] 119  16
```

Update the categorical variables with numberical variables

```
salesTemp = prop.table(table(training$sales_channel__c, training$target),1)
typeTemp = prop.table(table(training$type, training$target),1)
collectionTemp = prop.table(table(training$current_collection_method,
training$target),1)

data$sales1[data$sales_channel__c =="Direct"] <- 1-salesTemp[1]
data$sales1[data$sales_channel__c =="FAP: Managed Application Program"] <- 1-
salesTemp[2]
data$sales1[data$sales_channel__c =="Promontory"] <- 1-salesTemp[3]
data$sales1[data$sales_channel__c =="Referral"] <- 1-salesTemp[4]

data$type1[data$type =="Loan - New Customer"] <- 1-typeTemp[1]
data$type1[data$type =="Loan - Renewal"] <- 1-typeTemp[2]

data$collection1[data$current_collection_method =="ACH Pull"] <- 1-
collectionTemp[1]
data$collection1[data$current_collection_method =="Split Funding"] <- 1-
collectionTemp[2]
data$collection1[data$current_collection_method =="Transfer Account Vendors"]
<- 1-collectionTemp[3]

str(data)
```

```
## 'data.frame':    477 obs. of  19 variables:
##  $ as_of_date                      : Factor w/ 1 level "11/1/12": 1 1 1 1
1 1 1 1 1 1 ...
##  $ days_delinquent_old             : int  180 9 56 19 35 12 180 1 1 46
...
##  $ days_delinquent_new             : int  180 9 75 30 54 26 180 6 0 65
...
##  $ new_outstanding_principal_balance: num  29384 3200 56207 47496 21012
...
##  $ initial_loan_amount             : int  50000 10000 60000 50000 25000
20000 25000 10000 10000 15000 ...
##  $ fico                            : num  641 631 671 626 587 706 641 654
593 537 ...
##  $ sales_channel__c                : Factor w/ 4 levels "Direct","FAP:
Managed Application Program",..: 2 4 2 1 2 2 2 2 2 2 ...
##  $ type                            : Factor w/ 2 levels "Loan - New
Customer",..: 1 2 2 1 2 1 2 1 1 1 ...
##  $ current_collection_method       : Factor w/ 3 levels "ACH Pull","Split
Funding",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ term                            : int  12 6 9 6 6 12 6 6 6 6 ...
##  $ lender_payoff                   : num  0 0 0 0 0 ...
```

```
##  $ average_bank_balance__c          : num  46445 1284 29416 61028 2046 ...
##  $ last_cleared_payment_date        : Factor w/ 80 levels
"10/1/12","10/11/12",..: 22 22 42 19 80 22 67 18 18 54 ...
##  $ diff                             : int  0 0 19 11 19 14 0 5 -1 19 ...
##  $ target                           : num  0 0 1 1 1 1 0 1 0 1 ...
##  $ lender1                          : num  0 0 0 0 0 0 0 0 0 1 ...
##  $ sales1                           : num  0.658 0.56 0.658 0.722 0.658
...
##  $ type1                            : num  0.696 0.558 0.558 0.696 0.558
...
##  $ collection1                      : num  0.652 0.652 0.652 0.652 0.652
...
```

Exploratory data analysis

```
# "clients that get worse" vs "clients that remain the same group level or
get better".
table(data$target)

##
##   0   1
## 163 314

prop.table(table(data$target))

##
##         0         1
## 0.3417191 0.6582809

#explortory analysis for binary variables
#sales_channel__c, 4 level
table(data$sales_channel__c, data$target)

##
##                                    0   1
##   Direct                          19  55
##   FAP: Managed Application Program 129 241
##   Promontory                       1   0
##   Referral                        14  18

prop.table(table(data$sales_channel__c, data$target),1)

##
##                                           0         1
##   Direct                          0.2567568 0.7432432
##   FAP: Managed Application Program 0.3486486 0.6513514
##   Promontory                      1.0000000 0.0000000
##   Referral                        0.4375000 0.5625000

#type, 2 level
table(data$type, data$target)
```

```
##
##                          0   1
##   Loan - New Customer 109 243
##   Loan - Renewal       54  71
```

```
prop.table(table(data$type, data$target),1)
```

```
##
##                                0         1
##   Loan - New Customer 0.3096591 0.6903409
##   Loan - Renewal      0.4320000 0.5680000
```

```
#current_collection_method, 3 level
table(data$current_collection_method, data$target)
```

```
##
##                               0   1
##   ACH Pull                  145 278
##   Split Funding               3  32
##   Transfer Account Vendors  15   4
```

```
prop.table(table(data$current_collection_method, data$target),1)
```

```
##
##                                     0          1
##   ACH Pull                  0.34278960 0.65721040
##   Split Funding             0.08571429 0.91428571
##   Transfer Account Vendors 0.78947368 0.21052632
```

```
#term,
table(data$term, data$target)
```

```
##
##        0   1
##   3    1   0
##   4    0   1
##   5    0   2
##   6   94 219
##   9   20  19
##   12  30  53
##   15   1   3
##   18  17  17
```

```
prop.table(table(data$term, data$target),1)
```

```
##
##              0         1
##   3  1.0000000 0.0000000
##   4  0.0000000 1.0000000
##   5  0.0000000 1.0000000
##   6  0.3003195 0.6996805
##   9  0.5128205 0.4871795
##   12 0.3614458 0.6385542
```

```
##   15 0.2500000 0.7500000
##   18 0.5000000 0.5000000
```

```
#lender_payoff
data$lender1[data$lender_payoff == 0] <- 0
data$lender1[data$lender_payoff > 0] <- 1
```

```
table(data$lender1, data$target)
```

```
##
##       0   1
##   0 147 273
##   1  16  41
```

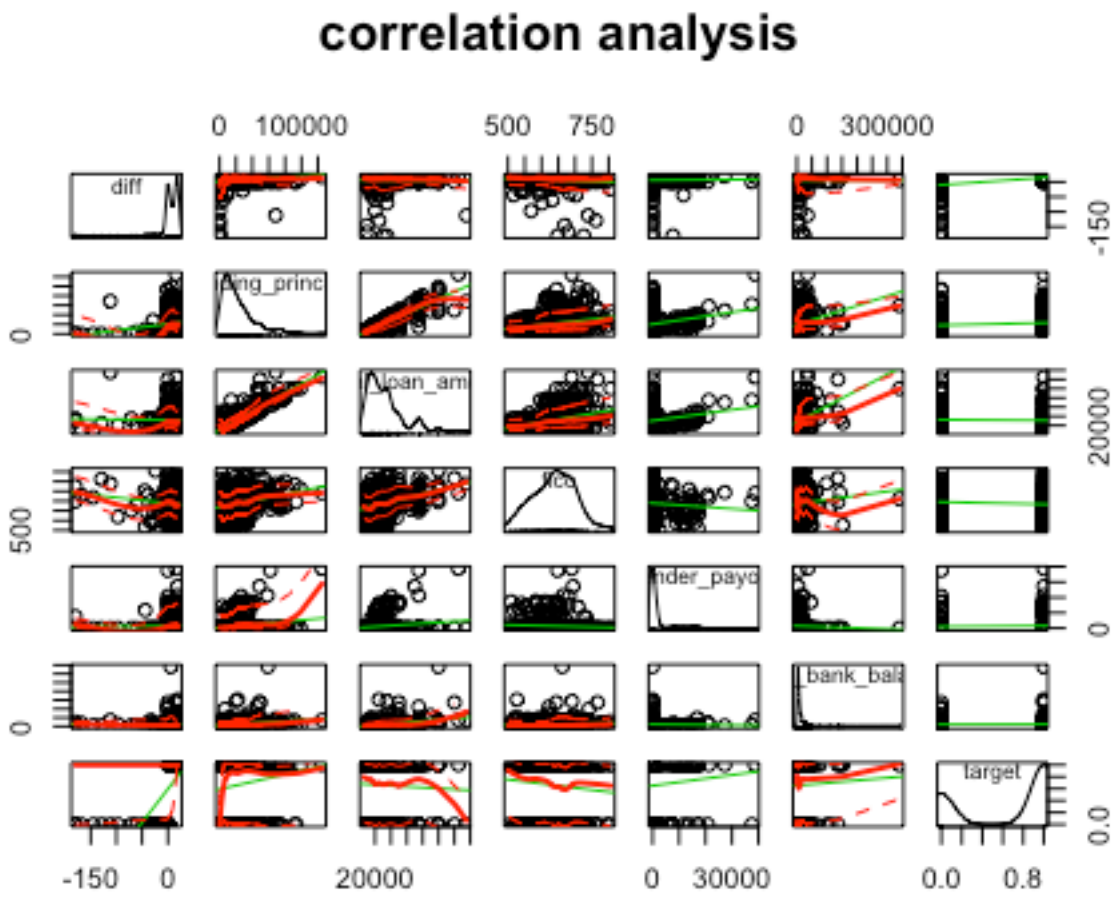
```
prop.table(table(data$lender1, data$target),1)
```

```
##
##             0         1
##   0 0.3500000 0.6500000
##   1 0.2807018 0.7192982
```

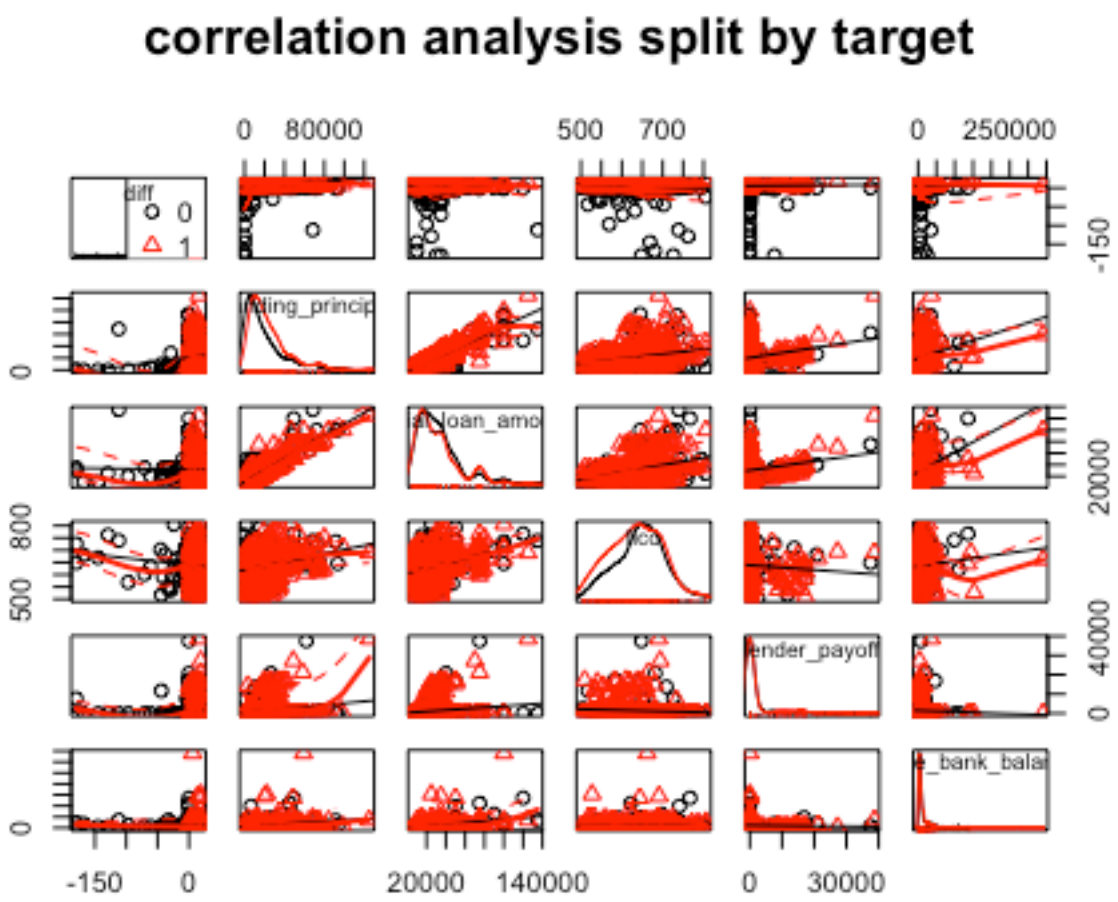
```
#correlation between the continues variables, include lowess and linear best fit lines, and boxplot, densities, or histograms in the principal diagonal, as well as rug plots in the margins of the cells.
```

```
dataConti =
data[c("diff","target","new_outstanding_principal_balance","initial_loan_amount","fico","lender_payoff","average_bank_balance__c")]
```

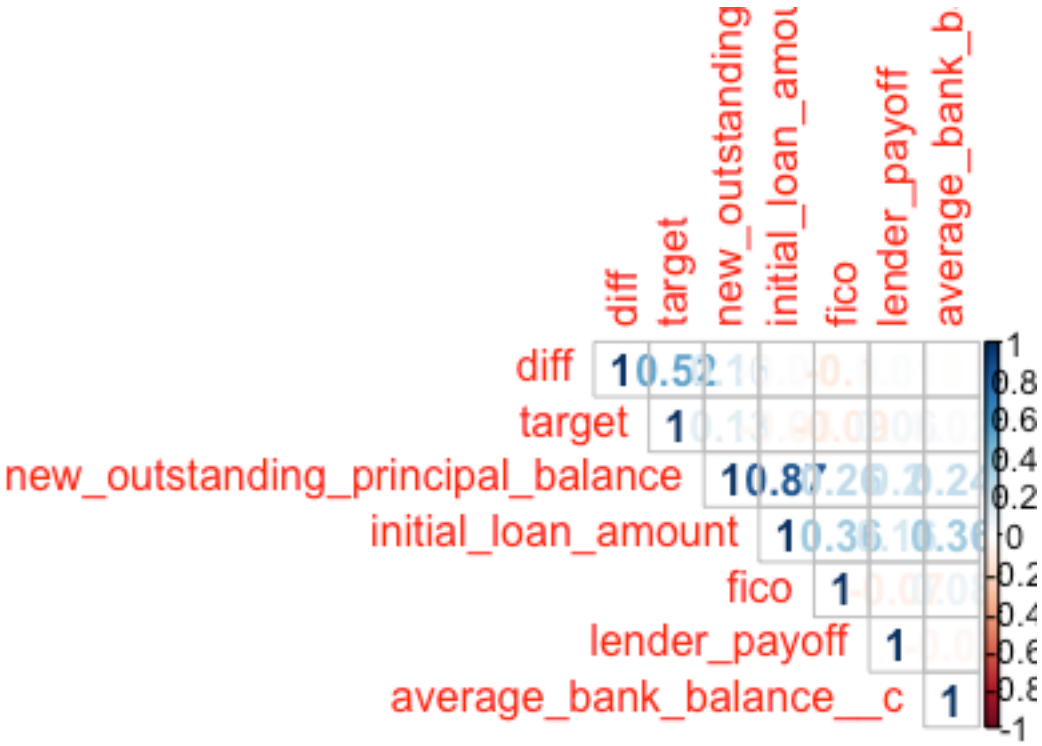
```
scatterplotMatrix(~diff+new_outstanding_principal_balance+initial_loan_amount+fico+lender_payoff+average_bank_balance__c+target, data = dataConti, main = "correlation analysis")
```




```
scatterplotMatrix(~diff+new_outstanding_principal_balance+initial_loan_amount
+fico+lender_payoff+average_bank_balance__c+target|target, data = dataConti,
main = "correlation analysis split by target")
```




```
M <- cor(dataConti)
corrplot(M,method = "number",type="upper")
```


diff
target
new_outstanding
initial_loan_amou
fico
lender_payoff
average_bank_b
diff
target
new_outstanding_principal_balance
initial_loan_amount
fico
lender_payoff
average_bank_balance__c
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1