DATASET

Stepan

20 09 2020

# Download data

#Set Working Directory  
setwd('C:/Users/Stepan/Desktop/6 курс/Машинное обучение')  
#OR Choose your Directory in 'Files' and click on 'More' -> 'Set as Working Directory'   
#Download file to the table. Source file is 'flats.csv'  
f <- read.csv2('DATASET.csv', header = TRUE, encoding = 'UNICOD')  
#Connect library  
library (dplyr)

## Warning: package 'dplyr' was built under R version 3.6.3

##   
## 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

#Have a look at the data  
glimpse(f)

## Rows: 17,170  
## Columns: 31  
## $ LOAN\_PRODUCT\_TYPE <fct> Cпоживчий...  
## $ LOAN\_AMOUNT <dbl> 6224.07, ...  
## $ BRANCH\_REGION <fct> Харків, Х...  
## $ CLIENT\_GENDER <fct> Жінка, Жі...  
## $ CLIENT\_FAMILYSTATUS <fct> В громадя...  
## $ CLIENT\_COUNTDEPENDENTS <int> 0, 0, 0, ...  
## $ CLIENT\_TOTALEXPERIENCE <dbl> 31, 3, 20...  
## $ CLIENT\_LASTEXPERIENCE <dbl> 12.0, 0.6...  
## $ CLIENT\_EDUCATION <fct> Середня -...  
## $ CLIENT\_ACTIVITYTYPE <fct> інша, тор...  
## $ CLIENT\_TOGETHER\_INCOME <dbl> 3750.00, ...  
## $ LOAN\_OUTSTANDINGLOANSCOUNT <int> 2, 5, 0, ...  
## $ LOAN\_EXISTINGCUSTOMERFLAG <fct> Ні, Ні, Н...  
## $ LOAN\_OVERDUE\_EXIST\_FLAG <fct> Ні, Так, ...  
## $ SCR\_WORKPLACECONFIRMED <fct> Не визнач...  
## $ EMPLOYMENTTYPE <fct> Офіційно ...  
## $ DECADE <fct> вторая де...  
## $ LOAN\_TERM <int> 364, 1095...  
## $ ZODIAC <fct> рак, весы...  
## $ ZODIAC\_CHINA <fct> Свинья, С...  
## $ AGE <int> 45, 22, 4...  
## $ CHANGE\_WORK <fct> була змі...  
## $ MATCH\_OF\_THE\_REGISTRATION\_PLACE\_WITH\_THE\_ACTUAL\_RESIDENCE <fct> ИСТИНА, И...  
## $ FAMILY\_PHONE <fct> не вказан...  
## $ REAL\_ESTATE <fct> не має вл...  
## $ CAR <fct> не має вл...  
## $ INFORMATION\_IN\_MGB <fct> відсутня,...  
## $ INFORMATION\_IN\_BKI <fct> наявна, в...  
## $ STRABIS <fct> не має оз...  
## $ PERCENT\_IN\_THE\_LOAN\_AMOUNT <dbl> 1.66, 3.7...  
## $ DELAY <fct> нет прост...

head(f)

## LOAN\_PRODUCT\_TYPE LOAN\_AMOUNT BRANCH\_REGION  
## 1 Cпоживчий кредит без довiдки про доходи 6224.07 Харків  
## 2 Cпоживчий кредит без довiдки про доходи 16816.14 Харків  
## 3 Cпоживчий кредит без довiдки про доходи 15560.17 Полтавська область  
## 4 Cпоживчий кредит без довiдки про доходи 5605.38 Харків  
## 5 Cпоживчий кредит без довiдки про доходи 16816.14 Харків  
## 6 Cпоживчий кредит без довiдки про доходи 10775.86 Полтавська область  
## CLIENT\_GENDER CLIENT\_FAMILYSTATUS CLIENT\_COUNTDEPENDENTS  
## 1 Жінка В громадянскому шлюбі 0  
## 2 Жінка Неодружений(-на) 0  
## 3 Жінка Розлучений (-на) 0  
## 4 Чоловік Одружений (-на) 0  
## 5 Чоловік Одружений (-на) 0  
## 6 Жінка Вдова (-ець) 0  
## CLIENT\_TOTALEXPERIENCE CLIENT\_LASTEXPERIENCE CLIENT\_EDUCATION  
## 1 31 12.0 Середня - технічна  
## 2 3 0.6 Середня  
## 3 20 12.0 Вища освіта  
## 4 29 12.5 Середня - технічна  
## 5 30 2.0 Середня - технічна  
## 6 30 23.0 Середня - технічна  
## CLIENT\_ACTIVITYTYPE CLIENT\_TOGETHER\_INCOME LOAN\_OUTSTANDINGLOANSCOUNT  
## 1 інша 3750 2  
## 2 торгівля та послуги 4450 5  
## 3 торгівля та послуги 8000 0  
## 4 інша 2150 1  
## 5 торгівля та послуги 6000 2  
## 6 інша 3500 1  
## LOAN\_EXISTINGCUSTOMERFLAG LOAN\_OVERDUE\_EXIST\_FLAG SCR\_WORKPLACECONFIRMED  
## 1 Ні Ні Не визначено  
## 2 Ні Так Не визначено  
## 3 Ні Ні Не визначено  
## 4 Ні Так Не визначено  
## 5 Так Ні Не визначено  
## 6 Ні Так Не визначено  
## EMPLOYMENTTYPE DECADE LOAN\_TERM ZODIAC ZODIAC\_CHINA AGE  
## 1 Офіційно працевлаштований вторая декада 364 рак Свинья 45  
## 2 Офіційно працевлаштований вторая декада 1095 весы Собака 22  
## 3 ФОП третья декада 364 телец Дракон 41  
## 4 Не працює третья декада 1095 телец Свинья 45  
## 5 Офіційно працевлаштований третья декада 1095 лев Тигр 54  
## 6 Офіційно працевлаштований третья декада 729 рак Коза 49  
## CHANGE\_WORK  
## 1 була зміна місця роботи  
## 2 була зміна місця роботи  
## 3 була зміна місця роботи  
## 4 була зміна місця роботи  
## 5 була зміна місця роботи  
## 6 була зміна місця роботи  
## MATCH\_OF\_THE\_REGISTRATION\_PLACE\_WITH\_THE\_ACTUAL\_RESIDENCE  
## 1 ИСТИНА  
## 2 ИСТИНА  
## 3 ИСТИНА  
## 4 ИСТИНА  
## 5 ИСТИНА  
## 6 ИСТИНА  
## FAMILY\_PHONE REAL\_ESTATE CAR  
## 1 не вказано телефон рідних не має власної нерухомості не має власного  
## 2 вказано телефон рідних не має власної нерухомості не має власного  
## 3 не вказано телефон рідних має власну нерухомість не має власного  
## 4 не вказано телефон рідних має власну нерухомість не має власного  
## 5 не вказано телефон рідних має власну нерухомість не має власного  
## 6 не вказано телефон рідних має власну нерухомість не має власного  
## INFORMATION\_IN\_MGB INFORMATION\_IN\_BKI STRABIS  
## 1 відсутня наявна не має ознаки страбісу  
## 2 відсутня відсутня не має ознаки страбісу  
## 3 відсутня відсутня не має ознаки страбісу  
## 4 відсутня наявна не має ознаки страбісу  
## 5 відсутня наявна не має ознаки страбісу  
## 6 відсутня наявна не має ознаки страбісу  
## PERCENT\_IN\_THE\_LOAN\_AMOUNT DELAY  
## 1 1.66 нет прострочки  
## 2 3.78 нет прострочки  
## 3 1.95 нет прострочки  
## 4 2.61 есть прострочка  
## 5 2.80 нет прострочки  
## 6 3.08 есть прострочка

# Висновок: Кількість спостережень - 17170, число змінних - 31

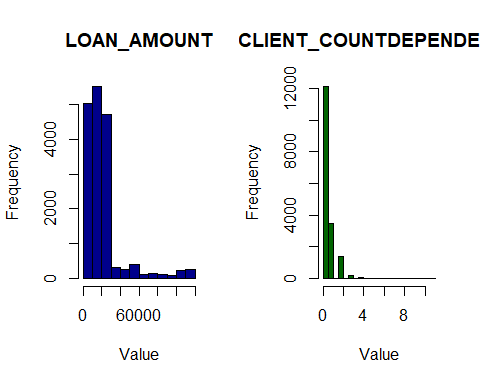
# Visualising

## Histogram

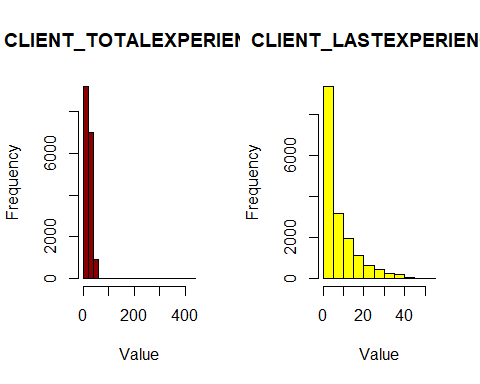
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.6.3

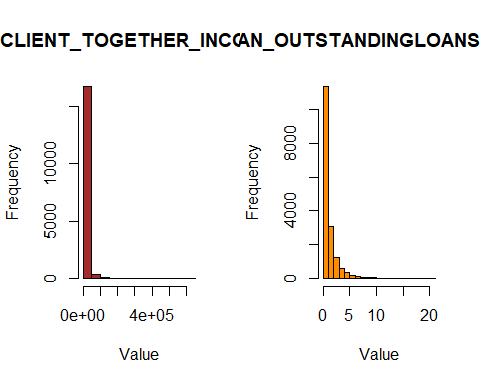
par(mfrow = c(1, 2))  
hist(f$LOAN\_AMOUNT, col = 'dark blue', main = 'LOAN\_AMOUNT', xlab = 'Value')  
hist(f$CLIENT\_COUNTDEPENDENTS, col = 'dark green', main = 'CLIENT\_COUNTDEPENDENTS', xlab = 'Value')



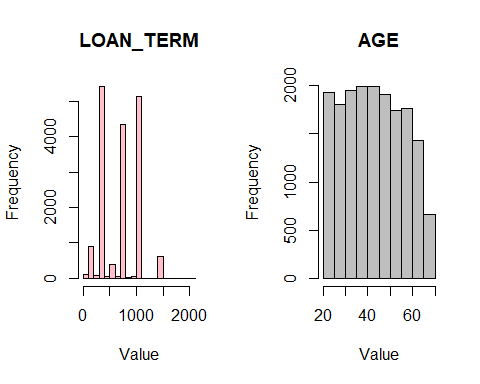
hist(f$CLIENT\_TOTALEXPERIENCE, col = 'dark red', main = 'CLIENT\_TOTALEXPERIENCE', xlab = 'Value')  
hist(f$CLIENT\_LASTEXPERIENCE, col = 'yellow', main = 'CLIENT\_LASTEXPERIENCE', xlab = 'Value')



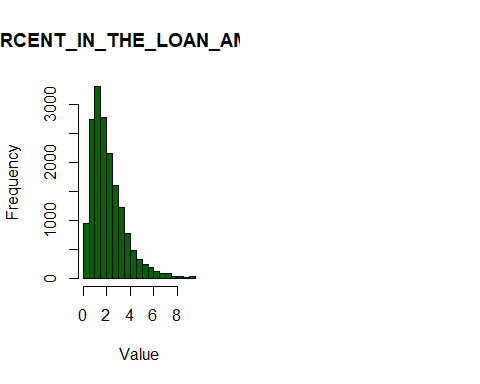
hist(f$CLIENT\_TOGETHER\_INCOME, col = 'brown', main = 'CLIENT\_TOGETHER\_INCOME', xlab = 'Value')  
hist(f$LOAN\_OUTSTANDINGLOANSCOUNT, col = 'dark orange', main = 'LOAN\_OUTSTANDINGLOANSCOUNT', xlab = 'Value')



hist(f$LOAN\_TERM, col = 'pink', main = 'LOAN\_TERM', xlab = 'Value')  
hist(f$AGE, col = 'gray', main = 'AGE', xlab = 'Value')



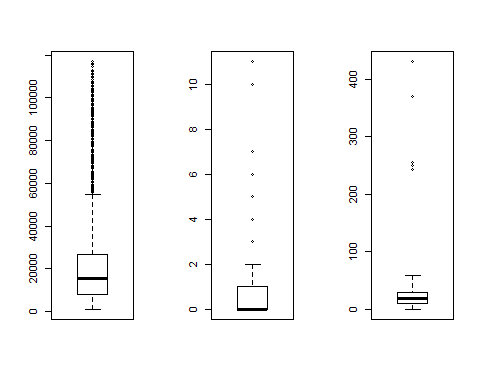
hist(f$PERCENT\_IN\_THE\_LOAN\_AMOUNT, col = 'dark green', main = 'PERCENT\_IN\_THE\_LOAN\_AMOUNT', xlab = 'Value')



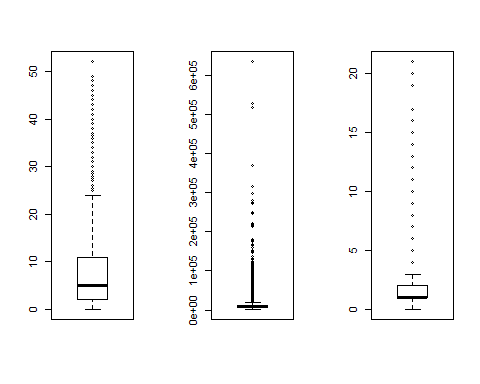
**# Висновок: розподіл змінної LOAN\_AMOUNT та PERCENT\_IN\_THE\_LOAN\_AMOUNT мають довгий хвіст.**

**# Розподіл змінної AGE має рівномірний розподіл. # Розподіл змінної LOAN-TERM має нетиповий вигляд, з огляду на те що договора кредитування мають стандартні терміни повернення коштів. ## Box-plot**

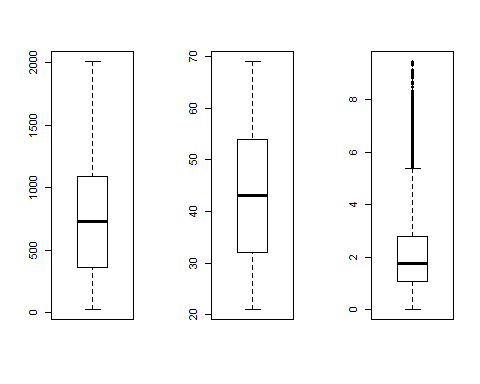
par(mfrow = c(1, 3))  
boxplot(f$LOAN\_AMOUNT)  
boxplot(f$CLIENT\_COUNTDEPENDENTS)  
boxplot(f$CLIENT\_TOTALEXPERIENCE)



boxplot(f$CLIENT\_LASTEXPERIENCE)  
boxplot(f$CLIENT\_TOGETHER\_INCOME)  
boxplot(f$LOAN\_OUTSTANDINGLOANSCOUNT)



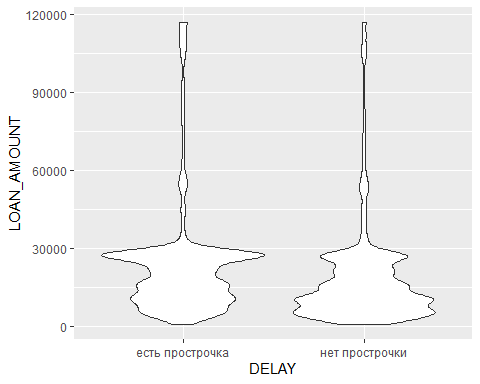
boxplot(f$LOAN\_TERM)  
boxplot(f$AGE)  
boxplot(f$PERCENT\_IN\_THE\_LOAN\_AMOUNT)



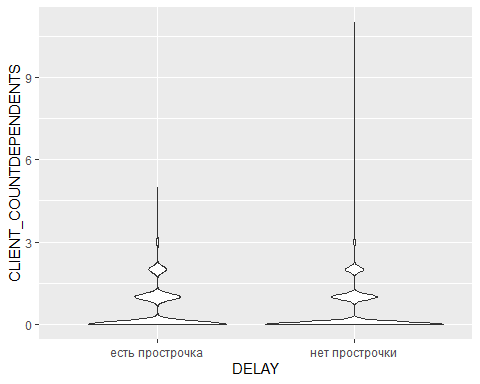
#qplot(data = f,   
 #x = condition,   
 # y = price,   
 #geom = "boxplot")

**#Висновок: всі змінні m2 та price мають викиди. ## Violin**

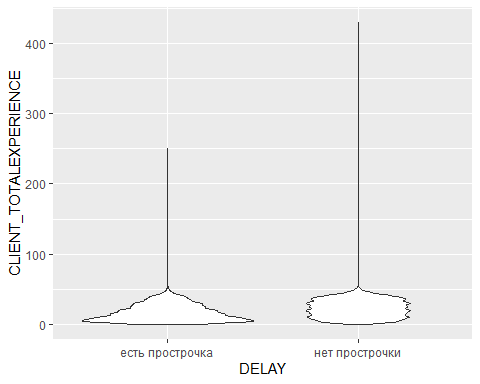
qplot(data = f,   
 x = DELAY,   
 y = LOAN\_AMOUNT,   
 geom = "violin")



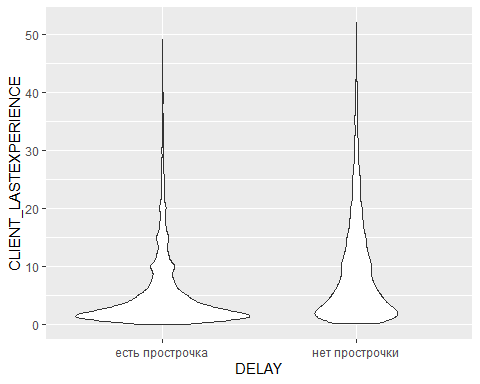
qplot(data = f,   
 x = DELAY,   
 y = CLIENT\_COUNTDEPENDENTS,   
 geom = "violin")



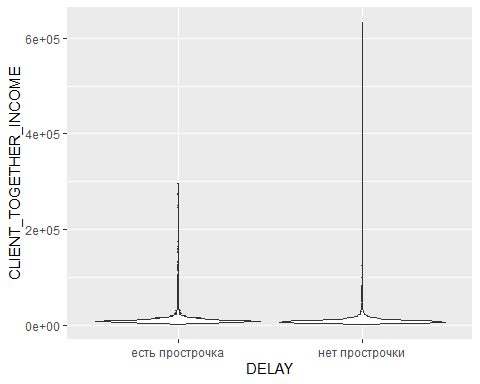
qplot(data = f,   
 x = DELAY,   
 y = CLIENT\_TOTALEXPERIENCE,   
 geom = "violin")



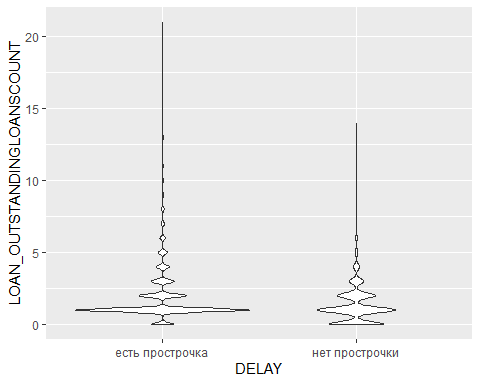
qplot(data = f,   
 x = DELAY,   
 y = CLIENT\_LASTEXPERIENCE,   
 geom = "violin")



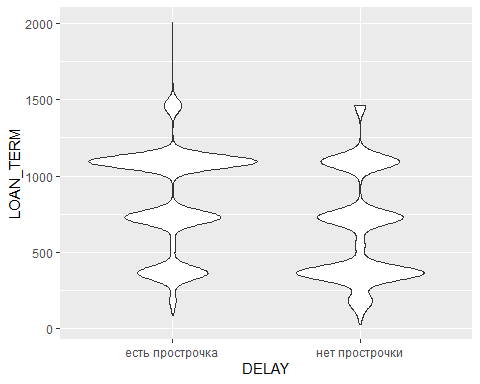
qplot(data = f,   
 x = DELAY,   
 y = CLIENT\_TOGETHER\_INCOME,   
 geom = "violin")



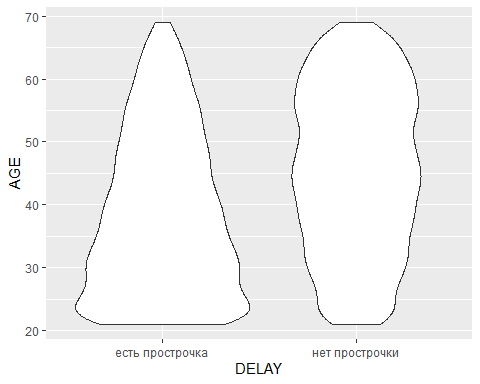
qplot(data = f,   
 x = DELAY,   
 y = LOAN\_OUTSTANDINGLOANSCOUNT,   
 geom = "violin")



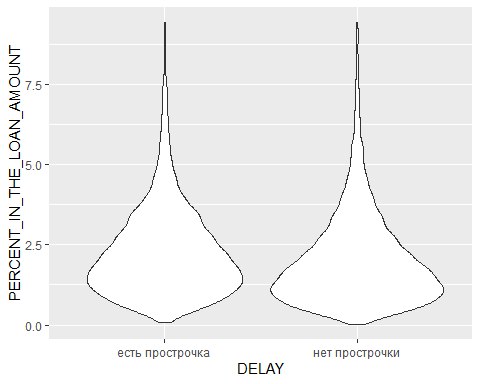
qplot(data = f,   
 x = DELAY,   
 y = LOAN\_TERM,   
 geom = "violin")



qplot(data = f,   
 x = DELAY,   
 y = AGE,   
 geom = "violin")



qplot(data = f,   
 x = DELAY,   
 y = PERCENT\_IN\_THE\_LOAN\_AMOUNT,   
 geom = "violin")



**#Висновок: нетипові значеня змінної практично не залежать від наявності прострочення кредиту.**

# Statistics

## Descriptive statistics

library (psych)

## Warning: package 'psych' was built under R version 3.6.3

##   
## Attaching package: 'psych'

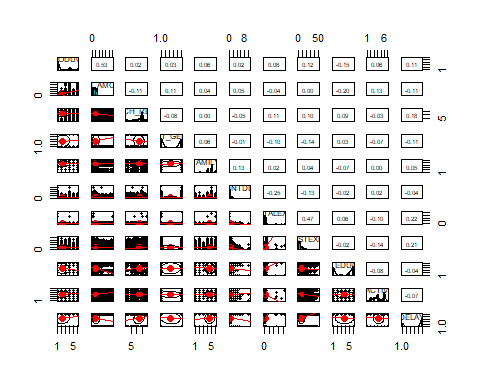
## The following objects are masked from 'package:ggplot2':  
##   
## %+%, alpha

describe(f)

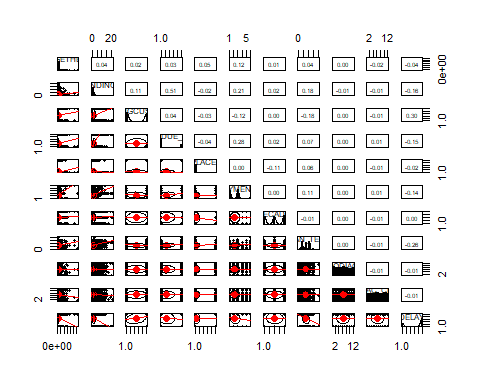
## vars n mean  
## LOAN\_PRODUCT\_TYPE\* 1 17170 2.17  
## LOAN\_AMOUNT 2 17170 20753.72  
## BRANCH\_REGION\* 3 17170 12.02  
## CLIENT\_GENDER\* 4 17170 1.46  
## CLIENT\_FAMILYSTATUS\* 5 17170 4.53  
## CLIENT\_COUNTDEPENDENTS 6 17170 0.40  
## CLIENT\_TOTALEXPERIENCE 7 17170 20.81  
## CLIENT\_LASTEXPERIENCE 8 17170 8.04  
## CLIENT\_EDUCATION\* 9 17170 3.78  
## CLIENT\_ACTIVITYTYPE\* 10 17170 4.21  
## CLIENT\_TOGETHER\_INCOME 11 17170 11249.65  
## LOAN\_OUTSTANDINGLOANSCOUNT 12 17170 1.51  
## LOAN\_EXISTINGCUSTOMERFLAG\* 13 17170 1.48  
## LOAN\_OVERDUE\_EXIST\_FLAG\* 14 17170 1.25  
## SCR\_WORKPLACECONFIRMED\* 15 17170 1.03  
## EMPLOYMENTTYPE\* 16 17170 2.25  
## DECADE\* 17 17170 2.03  
## LOAN\_TERM 18 17170 709.85  
## ZODIAC\* 19 17170 6.36  
## ZODIAC\_CHINA\* 20 17170 6.54  
## AGE 21 17170 43.15  
## CHANGE\_WORK\* 22 17170 1.14  
## MATCH\_OF\_THE\_REGISTRATION\_PLACE\_WITH\_THE\_ACTUAL\_RESIDENCE\* 23 17170 1.17  
## FAMILY\_PHONE\* 24 17170 1.58  
## REAL\_ESTATE\* 25 17170 1.34  
## CAR\* 26 17170 1.91  
## INFORMATION\_IN\_MGB\* 27 17170 1.39  
## INFORMATION\_IN\_BKI\* 28 17170 1.95  
## STRABIS\* 29 17170 1.81  
## PERCENT\_IN\_THE\_LOAN\_AMOUNT 30 17170 2.12  
## DELAY\* 31 17170 1.69  
## sd median  
## LOAN\_PRODUCT\_TYPE\* 1.77 1.00  
## LOAN\_AMOUNT 21247.81 15560.17  
## BRANCH\_REGION\* 4.72 14.00  
## CLIENT\_GENDER\* 0.50 1.00  
## CLIENT\_FAMILYSTATUS\* 1.20 5.00  
## CLIENT\_COUNTDEPENDENTS 0.71 0.00  
## CLIENT\_TOTALEXPERIENCE 13.33 20.00  
## CLIENT\_LASTEXPERIENCE 8.46 5.00  
## CLIENT\_EDUCATION\* 2.30 5.00  
## CLIENT\_ACTIVITYTYPE\* 1.58 4.00  
## CLIENT\_TOGETHER\_INCOME 17242.89 8000.00  
## LOAN\_OUTSTANDINGLOANSCOUNT 1.64 1.00  
## LOAN\_EXISTINGCUSTOMERFLAG\* 0.50 1.00  
## LOAN\_OVERDUE\_EXIST\_FLAG\* 0.43 1.00  
## SCR\_WORKPLACECONFIRMED\* 0.18 1.00  
## EMPLOYMENTTYPE\* 1.26 3.00  
## DECADE\* 0.84 2.00  
## LOAN\_TERM 349.76 729.00  
## ZODIAC\* 3.45 6.00  
## ZODIAC\_CHINA\* 3.46 7.00  
## AGE 13.29 43.00  
## CHANGE\_WORK\* 0.35 1.00  
## MATCH\_OF\_THE\_REGISTRATION\_PLACE\_WITH\_THE\_ACTUAL\_RESIDENCE\* 0.37 1.00  
## FAMILY\_PHONE\* 0.49 2.00  
## REAL\_ESTATE\* 0.47 1.00  
## CAR\* 0.29 2.00  
## INFORMATION\_IN\_MGB\* 0.49 1.00  
## INFORMATION\_IN\_BKI\* 0.21 2.00  
## STRABIS\* 0.39 2.00  
## PERCENT\_IN\_THE\_LOAN\_AMOUNT 1.45 1.77  
## DELAY\* 0.46 2.00  
## trimmed mad  
## LOAN\_PRODUCT\_TYPE\* 1.83 0.00  
## LOAN\_AMOUNT 16363.27 13483.73  
## BRANCH\_REGION\* 12.50 2.97  
## CLIENT\_GENDER\* 1.45 0.00  
## CLIENT\_FAMILYSTATUS\* 4.71 1.48  
## CLIENT\_COUNTDEPENDENTS 0.24 0.00  
## CLIENT\_TOTALEXPERIENCE 20.30 14.83  
## CLIENT\_LASTEXPERIENCE 6.44 5.93  
## CLIENT\_EDUCATION\* 3.85 1.48  
## CLIENT\_ACTIVITYTYPE\* 4.28 2.97  
## CLIENT\_TOGETHER\_INCOME 8461.61 3706.50  
## LOAN\_OUTSTANDINGLOANSCOUNT 1.23 1.48  
## LOAN\_EXISTINGCUSTOMERFLAG\* 1.48 0.00  
## LOAN\_OVERDUE\_EXIST\_FLAG\* 1.18 0.00  
## SCR\_WORKPLACECONFIRMED\* 1.00 0.00  
## EMPLOYMENTTYPE\* 2.14 1.48  
## DECADE\* 2.03 1.48  
## LOAN\_TERM 703.80 541.15  
## ZODIAC\* 6.34 4.45  
## ZODIAC\_CHINA\* 6.56 4.45  
## AGE 42.99 16.31  
## CHANGE\_WORK\* 1.05 0.00  
## MATCH\_OF\_THE\_REGISTRATION\_PLACE\_WITH\_THE\_ACTUAL\_RESIDENCE\* 1.08 0.00  
## FAMILY\_PHONE\* 1.60 0.00  
## REAL\_ESTATE\* 1.30 0.00  
## CAR\* 2.00 0.00  
## INFORMATION\_IN\_MGB\* 1.36 0.00  
## INFORMATION\_IN\_BKI\* 2.00 0.00  
## STRABIS\* 1.88 0.00  
## PERCENT\_IN\_THE\_LOAN\_AMOUNT 1.93 1.19  
## DELAY\* 1.74 0.00  
## min max  
## LOAN\_PRODUCT\_TYPE\* 1.00 6.00  
## LOAN\_AMOUNT 1000.00 116822.43  
## BRANCH\_REGION\* 1.00 19.00  
## CLIENT\_GENDER\* 1.00 2.00  
## CLIENT\_FAMILYSTATUS\* 1.00 6.00  
## CLIENT\_COUNTDEPENDENTS 0.00 11.00  
## CLIENT\_TOTALEXPERIENCE 0.10 430.00  
## CLIENT\_LASTEXPERIENCE 0.00 52.00  
## CLIENT\_EDUCATION\* 1.00 6.00  
## CLIENT\_ACTIVITYTYPE\* 1.00 7.00  
## CLIENT\_TOGETHER\_INCOME 1452.00 633622.00  
## LOAN\_OUTSTANDINGLOANSCOUNT 0.00 21.00  
## LOAN\_EXISTINGCUSTOMERFLAG\* 1.00 2.00  
## LOAN\_OVERDUE\_EXIST\_FLAG\* 1.00 2.00  
## SCR\_WORKPLACECONFIRMED\* 1.00 2.00  
## EMPLOYMENTTYPE\* 1.00 6.00  
## DECADE\* 1.00 3.00  
## LOAN\_TERM 28.00 2007.00  
## ZODIAC\* 1.00 12.00  
## ZODIAC\_CHINA\* 1.00 12.00  
## AGE 21.00 69.00  
## CHANGE\_WORK\* 1.00 2.00  
## MATCH\_OF\_THE\_REGISTRATION\_PLACE\_WITH\_THE\_ACTUAL\_RESIDENCE\* 1.00 2.00  
## FAMILY\_PHONE\* 1.00 2.00  
## REAL\_ESTATE\* 1.00 2.00  
## CAR\* 1.00 2.00  
## INFORMATION\_IN\_MGB\* 1.00 2.00  
## INFORMATION\_IN\_BKI\* 1.00 2.00  
## STRABIS\* 1.00 2.00  
## PERCENT\_IN\_THE\_LOAN\_AMOUNT 0.02 9.44  
## DELAY\* 1.00 2.00  
## range skew  
## LOAN\_PRODUCT\_TYPE\* 5.00 1.10  
## LOAN\_AMOUNT 115822.43 2.69  
## BRANCH\_REGION\* 18.00 -0.89  
## CLIENT\_GENDER\* 1.00 0.16  
## CLIENT\_FAMILYSTATUS\* 5.00 -1.34  
## CLIENT\_COUNTDEPENDENTS 11.00 2.29  
## CLIENT\_TOTALEXPERIENCE 429.90 3.79  
## CLIENT\_LASTEXPERIENCE 52.00 1.71  
## CLIENT\_EDUCATION\* 5.00 -0.28  
## CLIENT\_ACTIVITYTYPE\* 6.00 -0.15  
## CLIENT\_TOGETHER\_INCOME 632170.00 13.74  
## LOAN\_OUTSTANDINGLOANSCOUNT 21.00 3.06  
## LOAN\_EXISTINGCUSTOMERFLAG\* 1.00 0.07  
## LOAN\_OVERDUE\_EXIST\_FLAG\* 1.00 1.17  
## SCR\_WORKPLACECONFIRMED\* 1.00 5.31  
## EMPLOYMENTTYPE\* 5.00 0.80  
## DECADE\* 2.00 -0.05  
## LOAN\_TERM 1979.00 0.22  
## ZODIAC\* 11.00 0.04  
## ZODIAC\_CHINA\* 11.00 -0.03  
## AGE 48.00 0.08  
## CHANGE\_WORK\* 1.00 2.06  
## MATCH\_OF\_THE\_REGISTRATION\_PLACE\_WITH\_THE\_ACTUAL\_RESIDENCE\* 1.00 1.79  
## FAMILY\_PHONE\* 1.00 -0.31  
## REAL\_ESTATE\* 1.00 0.68  
## CAR\* 1.00 -2.79  
## INFORMATION\_IN\_MGB\* 1.00 0.46  
## INFORMATION\_IN\_BKI\* 1.00 -4.25  
## STRABIS\* 1.00 -1.56  
## PERCENT\_IN\_THE\_LOAN\_AMOUNT 9.42 1.48  
## DELAY\* 1.00 -0.82  
## kurtosis se  
## LOAN\_PRODUCT\_TYPE\* -0.36 0.01  
## LOAN\_AMOUNT 8.02 162.15  
## BRANCH\_REGION\* -0.12 0.04  
## CLIENT\_GENDER\* -1.98 0.00  
## CLIENT\_FAMILYSTATUS\* 1.59 0.01  
## CLIENT\_COUNTDEPENDENTS 10.54 0.01  
## CLIENT\_TOTALEXPERIENCE 92.81 0.10  
## CLIENT\_LASTEXPERIENCE 2.85 0.06  
## CLIENT\_EDUCATION\* -1.81 0.02  
## CLIENT\_ACTIVITYTYPE\* -1.18 0.01  
## CLIENT\_TOGETHER\_INCOME 345.64 131.59  
## LOAN\_OUTSTANDINGLOANSCOUNT 16.41 0.01  
## LOAN\_EXISTINGCUSTOMERFLAG\* -2.00 0.00  
## LOAN\_OVERDUE\_EXIST\_FLAG\* -0.63 0.00  
## SCR\_WORKPLACECONFIRMED\* 26.25 0.00  
## EMPLOYMENTTYPE\* 0.77 0.01  
## DECADE\* -1.56 0.01  
## LOAN\_TERM -1.11 2.67  
## ZODIAC\* -1.18 0.03  
## ZODIAC\_CHINA\* -1.21 0.03  
## AGE -1.10 0.10  
## CHANGE\_WORK\* 2.26 0.00  
## MATCH\_OF\_THE\_REGISTRATION\_PLACE\_WITH\_THE\_ACTUAL\_RESIDENCE\* 1.21 0.00  
## FAMILY\_PHONE\* -1.91 0.00  
## REAL\_ESTATE\* -1.53 0.00  
## CAR\* 5.76 0.00  
## INFORMATION\_IN\_MGB\* -1.79 0.00  
## INFORMATION\_IN\_BKI\* 16.09 0.00  
## STRABIS\* 0.44 0.00  
## PERCENT\_IN\_THE\_LOAN\_AMOUNT 2.89 0.01  
## DELAY\* -1.33 0.00

## Correlations

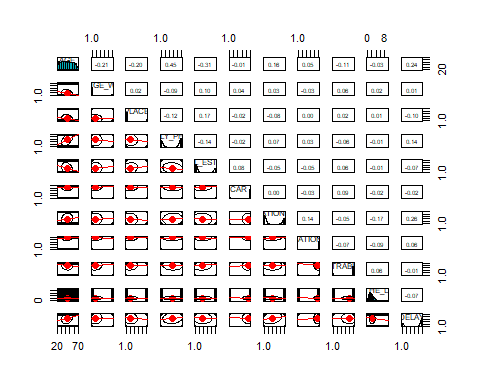
library(dplyr)  
library(psych)  
data\_s1 <- select(f, LOAN\_PRODUCT\_TYPE, LOAN\_AMOUNT, BRANCH\_REGION, CLIENT\_GENDER, CLIENT\_FAMILYSTATUS, CLIENT\_COUNTDEPENDENTS, CLIENT\_TOTALEXPERIENCE, CLIENT\_LASTEXPERIENCE, CLIENT\_EDUCATION, CLIENT\_ACTIVITYTYPE, DELAY)  
pairs.panels(data\_s1, lm=TRUE, # linear fit   
 method = "pearson", # correlation method  
 hist.col = "#00AFBB"  
 )



data\_s2 <- select(f, CLIENT\_TOGETHER\_INCOME, LOAN\_OUTSTANDINGLOANSCOUNT, LOAN\_EXISTINGCUSTOMERFLAG, LOAN\_OVERDUE\_EXIST\_FLAG, SCR\_WORKPLACECONFIRMED, EMPLOYMENTTYPE, DECADE, LOAN\_TERM, ZODIAC, ZODIAC\_CHINA, DELAY)  
pairs.panels(data\_s2, lm=TRUE, # linear fit   
 method = "pearson", # correlation method  
 hist.col = "#00AFBB"  
 )



data\_s3 <- select(f, AGE, CHANGE\_WORK, MATCH\_OF\_THE\_REGISTRATION\_PLACE\_WITH\_THE\_ACTUAL\_RESIDENCE, FAMILY\_PHONE, REAL\_ESTATE, CAR, INFORMATION\_IN\_MGB, INFORMATION\_IN\_BKI, STRABIS, PERCENT\_IN\_THE\_LOAN\_AMOUNT, DELAY)  
pairs.panels(data\_s3, lm=TRUE, # linear fit   
 method = "pearson", # correlation method  
 hist.col = "#00AFBB"  
 )



**#Висновок: найбільше на DELAY впливає змінна LOAN\_EXISTINGCUSTOMERFLAG. ## Remove the ejections (not recommended strategy)**

f\_ej1 <- f[f$LOAN\_AMOUNT < mean(f$LOAN\_AMOUNT)+sd(f$LOAN\_AMOUNT)\*3, ]  
describe(f\_ej1$LOAN\_AMOUNT)

## vars n mean sd median trimmed mad min max range  
## X1 1 16549 17549.52 13468.94 15000 15747.04 12242.21 1000 84112.15 83112.15  
## skew kurtosis se  
## X1 1.82 4.66 104.7

f\_ej2 <- f[f$CLIENT\_COUNTDEPENDENTS < mean(f$CLIENT\_COUNTDEPENDENTS)+sd(f$CLIENT\_COUNTDEPENDENTS)\*3, ]  
describe(f\_ej1$CLIENT\_COUNTDEPENDENTS)

## vars n mean sd median trimmed mad min max range skew kurtosis se  
## X1 1 16549 0.4 0.71 0 0.24 0 0 11 11 2.33 10.96 0.01

f\_ej3 <- f[f$CLIENT\_TOTALEXPERIENCE < mean(f$CLIENT\_TOTALEXPERIENCE)+sd(f$CLIENT\_TOTALEXPERIENCE)\*3, ]  
describe(f\_ej1$CLIENT\_TOTALEXPERIENCE)

## vars n mean sd median trimmed mad min max range skew kurtosis se  
## X1 1 16549 20.8 13.3 20 20.3 14.83 0.1 430 429.9 3.67 92.59 0.1

f\_ej4 <- f[f$CLIENT\_LASTEXPERIENCE < mean(f$CLIENT\_LASTEXPERIENCE)+sd(f$CLIENT\_LASTEXPERIENCE)\*3, ]  
describe(f\_ej1$CLIENT\_LASTEXPERIENCE)

## vars n mean sd median trimmed mad min max range skew kurtosis se  
## X1 1 16549 8.01 8.49 5 6.39 5.93 0 52 52 1.72 2.9 0.07

f\_ej5 <- f[f$CLIENT\_TOGETHER\_INCOME < mean(f$CLIENT\_TOGETHER\_INCOME)+sd(f$CLIENT\_TOGETHER\_INCOME)\*3, ]  
describe(f\_ej1$CLIENT\_TOGETHER\_INCOME)

## vars n mean sd median trimmed mad min max range  
## X1 1 16549 9927.29 12660.7 8000 8162.25 3409.98 1452 527657.5 526205.5  
## skew kurtosis se  
## X1 15.77 457 98.42

f\_ej6 <- f[f$LOAN\_OUTSTANDINGLOANSCOUNT < mean(f$LOAN\_OUTSTANDINGLOANSCOUNT)+sd(f$LOAN\_OUTSTANDINGLOANSCOUNT)\*3, ]  
describe(f\_ej1$LOAN\_OUTSTANDINGLOANSCOUNT)

## vars n mean sd median trimmed mad min max range skew kurtosis se  
## X1 1 16549 1.51 1.63 1 1.23 1.48 0 21 21 3.06 16.61 0.01

f\_ej7 <- f[f$LOAN\_TERM < mean(f$LOAN\_TERM)+sd(f$LOAN\_TERM)\*3, ]  
describe(f\_ej1$LOAN\_TERM)

## vars n mean sd median trimmed mad min max range skew kurtosis  
## X1 1 16549 697.17 341.57 729 694.14 541.15 28 2007 1979 0.2 -1.17  
## se  
## X1 2.66

f\_ej8 <- f[f$AGE < mean(f$AGE)+sd(f$AGE)\*3, ]  
describe(f\_ej1$AGE)

## vars n mean sd median trimmed mad min max range skew kurtosis se  
## X1 1 16549 43.21 13.36 43 43.06 16.31 21 69 48 0.07 -1.12 0.1

f\_ej9 <- f[f$PERCENT\_IN\_THE\_LOAN\_AMOUNT < mean(f$PERCENT\_IN\_THE\_LOAN\_AMOUNT)+sd(f$PERCENT\_IN\_THE\_LOAN\_AMOUNT)\*3, ]  
describe(f\_ej1$PERCENT\_IN\_THE\_LOAN\_AMOUNT)

## vars n mean sd median trimmed mad min max range skew kurtosis se  
## X1 1 16549 2.05 1.35 1.75 1.88 1.16 0.02 9.39 9.37 1.38 2.69 0.01

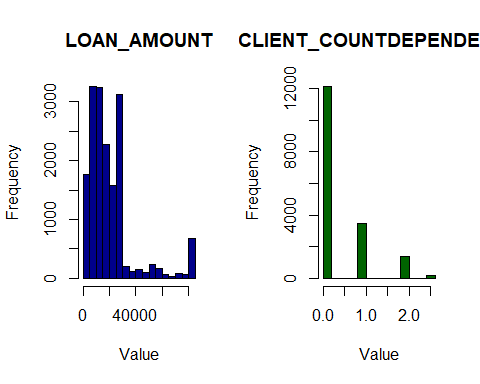
## Replace with max

f\_ej11 <- f  
f\_ej11$LOAN\_AMOUNT <- ifelse(f$LOAN\_AMOUNT < mean(f$LOAN\_AMOUNT)+sd(f$LOAN\_AMOUNT)\*3,f$LOAN\_AMOUNT,mean(f$LOAN\_AMOUNT)+sd(f$LOAN\_AMOUNT)\*3)  
#let's work with f\_eg2  
f <- f\_ej11  
  
f\_ej12 <- f  
f\_ej12$CLIENT\_COUNTDEPENDENTS <- ifelse(f$CLIENT\_COUNTDEPENDENTS < mean(f$CLIENT\_COUNTDEPENDENTS)+sd(f$CLIENT\_COUNTDEPENDENTS)\*3,f$CLIENT\_COUNTDEPENDENTS,mean(f$CLIENT\_COUNTDEPENDENTS)+sd(f$CLIENT\_COUNTDEPENDENTS)\*3)  
f <- f\_ej12  
  
f\_ej13 <- f  
f\_ej13$CLIENT\_TOTALEXPERIENCE <- ifelse(f$CLIENT\_TOTALEXPERIENCE < mean(f$CLIENT\_TOTALEXPERIENCE)+sd(f$CLIENT\_TOTALEXPERIENCE)\*3,f$CLIENT\_TOTALEXPERIENCE,mean(f$CLIENT\_TOTALEXPERIENCE)+sd(f$CLIENT\_TOTALEXPERIENCE)\*3)  
f <- f\_ej13  
  
f\_ej14 <- f  
f\_ej14$CLIENT\_LASTEXPERIENCE <- ifelse(f$CLIENT\_LASTEXPERIENCE < mean(f$CLIENT\_LASTEXPERIENCE)+sd(f$CLIENT\_LASTEXPERIENCE)\*3,f$CLIENT\_LASTEXPERIENCE,mean(f$CLIENT\_LASTEXPERIENCE)+sd(f$CLIENT\_LASTEXPERIENCE)\*3)  
f <- f\_ej14  
  
f\_ej15 <- f  
f\_ej15$CLIENT\_TOGETHER\_INCOME <- ifelse(f$CLIENT\_TOGETHER\_INCOME < mean(f$CLIENT\_TOGETHER\_INCOME)+sd(f$CLIENT\_TOGETHER\_INCOME)\*3,f$CLIENT\_TOGETHER\_INCOME,mean(f$CLIENT\_TOGETHER\_INCOME)+sd(f$CLIENT\_TOGETHER\_INCOME)\*3)  
f <- f\_ej15  
  
f\_ej16 <- f  
f\_ej16$LOAN\_OUTSTANDINGLOANSCOUNT <- ifelse(f$LOAN\_OUTSTANDINGLOANSCOUNT < mean(f$LOAN\_OUTSTANDINGLOANSCOUNT)+sd(f$LOAN\_OUTSTANDINGLOANSCOUNT)\*3,f$LOAN\_OUTSTANDINGLOANSCOUNT,mean(f$LOAN\_OUTSTANDINGLOANSCOUNT)+sd(f$LOAN\_OUTSTANDINGLOANSCOUNT)\*3)  
f <- f\_ej16  
  
f\_ej17 <- f  
f\_ej17$LOAN\_TERM <- ifelse(f$LOAN\_TERM < mean(f$LOAN\_TERM)+sd(f$LOAN\_TERM)\*3,f$LOAN\_TERM,mean(f$LOAN\_TERM)+sd(f$LOAN\_TERM)\*3)  
f <- f\_ej17  
  
f\_ej18 <- f  
f\_ej18$AGE <- ifelse(f$AGE < mean(f$AGE)+sd(f$AGE)\*3,f$AGE,mean(f$AGE)+sd(f$AGE)\*3)  
f <- f\_ej18  
  
f\_ej19 <- f  
f\_ej19$PERCENT\_IN\_THE\_LOAN\_AMOUNT <- ifelse(f$PERCENT\_IN\_THE\_LOAN\_AMOUNT < mean(f$PERCENT\_IN\_THE\_LOAN\_AMOUNT)+sd(f$PERCENT\_IN\_THE\_LOAN\_AMOUNT)\*3,f$PERCENT\_IN\_THE\_LOAN\_AMOUNT,mean(f$PERCENT\_IN\_THE\_LOAN\_AMOUNT)+sd(f$PERCENT\_IN\_THE\_LOAN\_AMOUNT)\*3)  
f <- f\_ej19  
  
describe(f\_ej12$PERCENT\_IN\_THE\_LOAN\_AMOUNT)

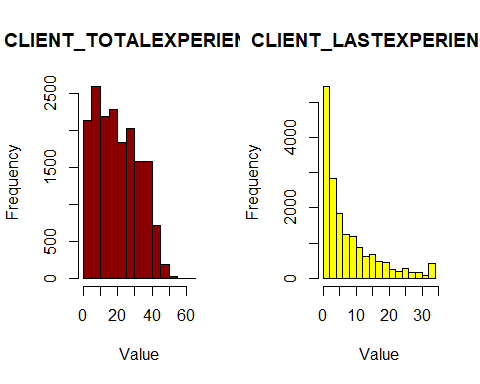
## vars n mean sd median trimmed mad min max range skew kurtosis se  
## X1 1 17170 2.12 1.45 1.77 1.93 1.19 0.02 9.44 9.42 1.48 2.89 0.01

**#Висновок: для корекції викидів обраний варіант заповнення граничними значеннями.**

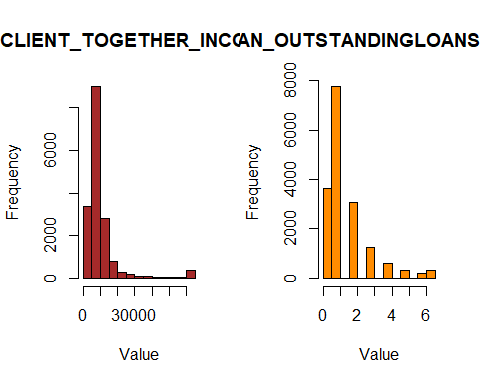
library(ggplot2)  
par(mfrow = c(1, 2))  
hist(f$LOAN\_AMOUNT, col = 'dark blue', main = 'LOAN\_AMOUNT', xlab = 'Value')  
hist(f$CLIENT\_COUNTDEPENDENTS, col = 'dark green', main = 'CLIENT\_COUNTDEPENDENTS', xlab = 'Value')



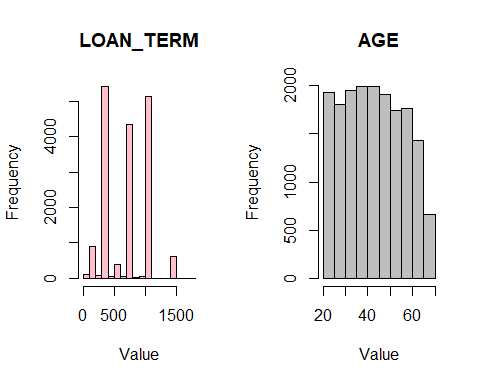
hist(f$CLIENT\_TOTALEXPERIENCE, col = 'dark red', main = 'CLIENT\_TOTALEXPERIENCE', xlab = 'Value')  
hist(f$CLIENT\_LASTEXPERIENCE, col = 'yellow', main = 'CLIENT\_LASTEXPERIENCE', xlab = 'Value')



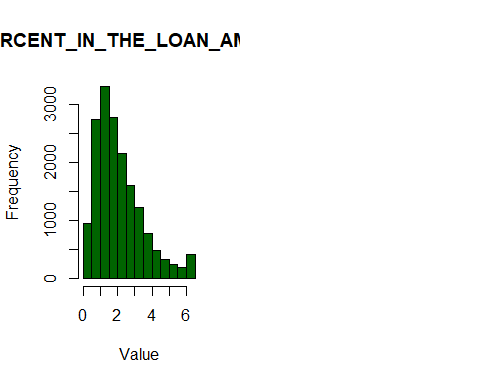
hist(f$CLIENT\_TOGETHER\_INCOME, col = 'brown', main = 'CLIENT\_TOGETHER\_INCOME', xlab = 'Value')  
hist(f$LOAN\_OUTSTANDINGLOANSCOUNT, col = 'dark orange', main = 'LOAN\_OUTSTANDINGLOANSCOUNT', xlab = 'Value')



hist(f$LOAN\_TERM, col = 'pink', main = 'LOAN\_TERM', xlab = 'Value')  
hist(f$AGE, col = 'gray', main = 'AGE', xlab = 'Value')



hist(f$PERCENT\_IN\_THE\_LOAN\_AMOUNT, col = 'dark green', main = 'PERCENT\_IN\_THE\_LOAN\_AMOUNT', xlab = 'Value')



# Splitting the dataset into the TRAIN set and TEST set

set.seed(123)  
library(caTools)

split = sample.split(f$DELAY, SplitRatio = 0.8)  
f\_train = subset(f, split == TRUE)  
f\_test = subset(f, split == FALSE)  
#Write prepared data to the file  
write.csv2(f\_train, file = "DATASET\_train.csv")  
write.csv2(f\_test, file = "DATASET\_test.csv")

**#Висновок: датасет розподілений на навчальну та тестову вибірки. Результати збережені в окремих файлах.**