Projet Visualisation des données

Projet final

Julien Muhlemann

6 June, 2024

Table of contents

Projet Visualisation des données (VID) **1**
Introduction 1
Description des données 2
Chargement des données 'GermanCredit.csv' tri en variable catégorielle et numérique 2
Affichage des données catégorielles et de leur fréquences. 2
Correction des valeurs aberrantes 17
Imputation des valeurs manquantes 17

Projet Visualisation des données (VID)

Introduction

Le jeu de données de crédit allemand comprend 1000 demandes de crédit passées, chacune décrite par 30 variables. L'objectif principal de cette analyse est d'identifier les caractéristiques pouvant déterminer la solvabilité des nouveaux demandeurs, classés comme des risques de crédit "Bons" ou "Mauvais". Ce rapport détaillera les caractéristiques des données, la méthodologie employée pour l'analyse et les conclusion sur le choix des variables expicatives les plus pertinentes pour la prédiction de la solvabilité des demandeurs.

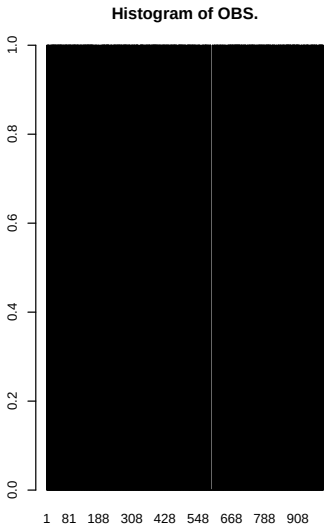
Description des données

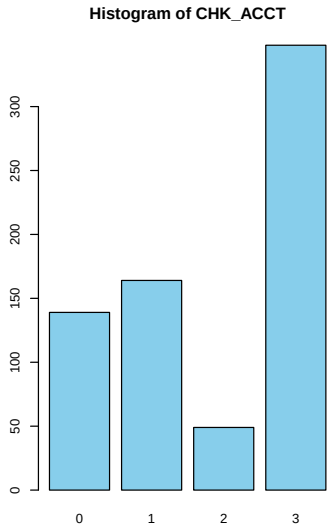
Chargement des données 'GermanCredit.csv' tri en variable catégorielle et numérique

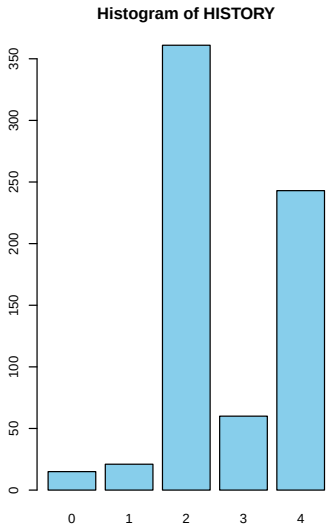
```r
german_credit <- read.csv("GermanCredit.csv", header=TRUE, sep=';')

quanti <- c(3, 11, 14, 23)
categorical_data <- german_credit[-quanti]
numeric_data <- german_credit[quanti]

for (col in names(categorical_data)) {
  categorical_data[[col]] <- factor(categorical_data[[col]])
}

for (col in names(numeric_data)) {
  numeric_data[[col]] <- as.numeric(numeric_data[[col]])
}
```

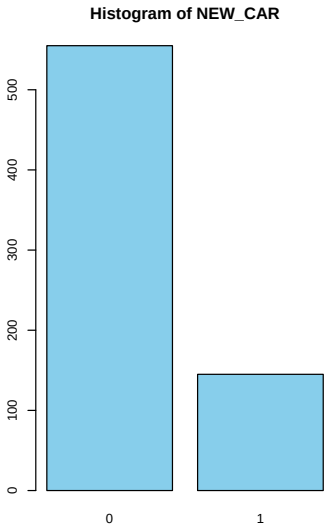
Affichage des données catégorielles et de leur fréquences.

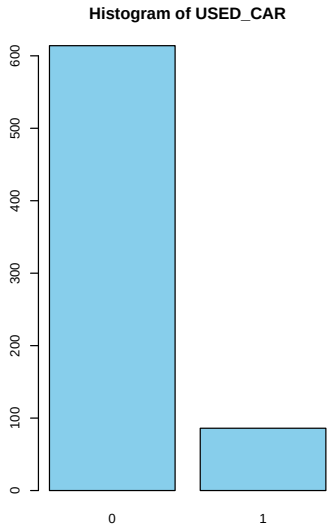
```r
# affichage des histogrammes pour les variables catégorielles
categorical_data_candidat = categorical_data[categorical_data$RESPONSE == 1,]

par(mfrow = c(2, 3))

for (col in names(categorical_data)) {
    barplot(table(categorical_data_candidat[[col]]), main = paste("Histogram of", col), co
}
```

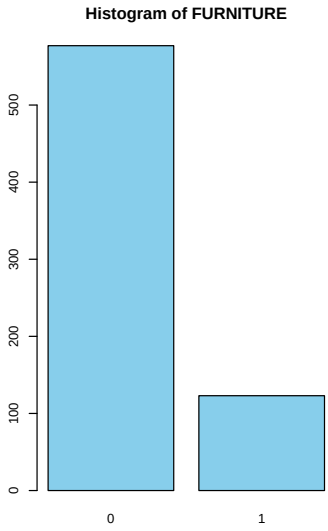

Histogram of OBS.
0.0
0.2
0.4
0.6
0.8
1.0
1
81
188
308
428
548
668
788
908

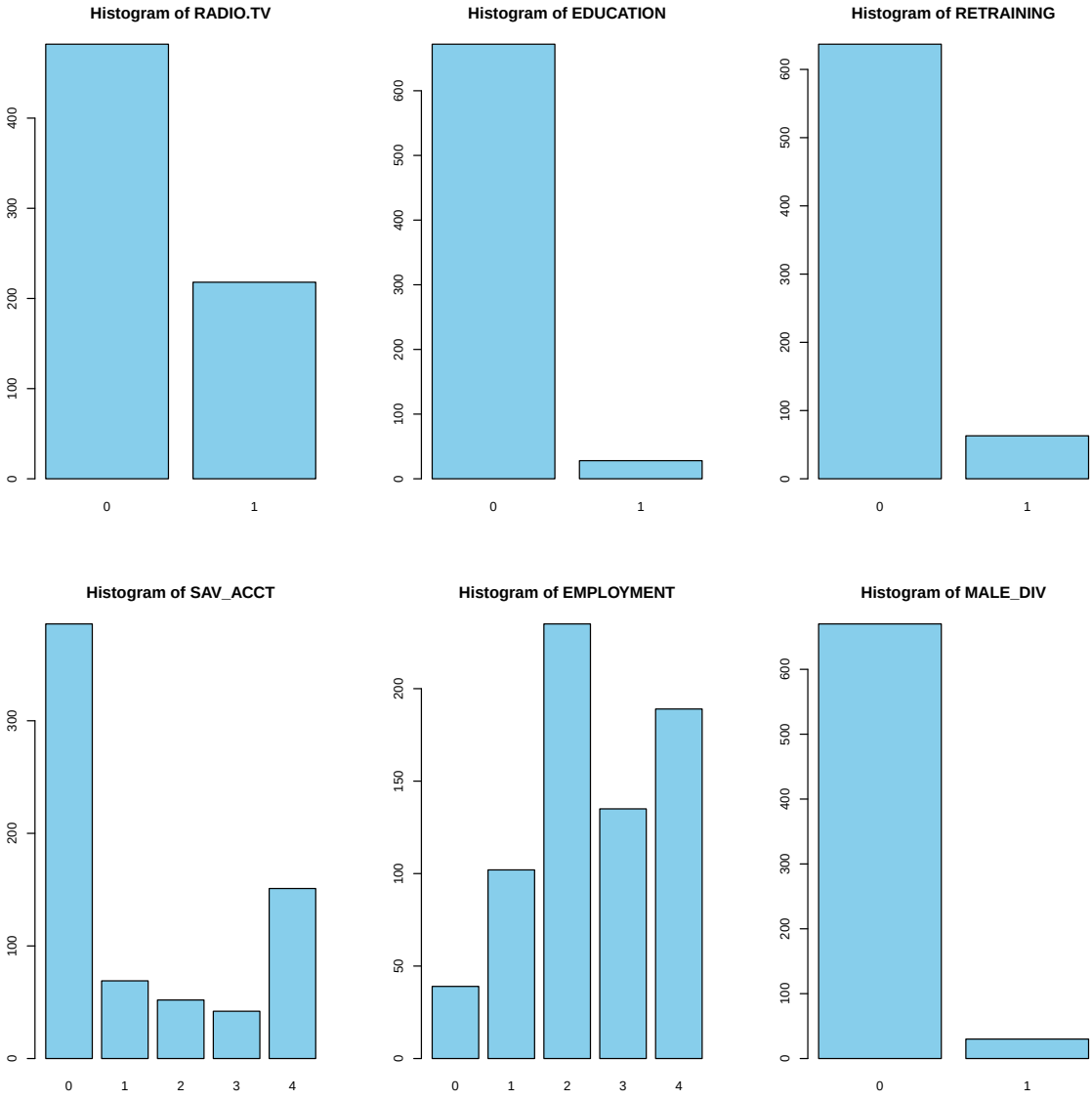

Histogram of CHK_ACCT
0
50
100
150
200
250
300
0
1
2
3

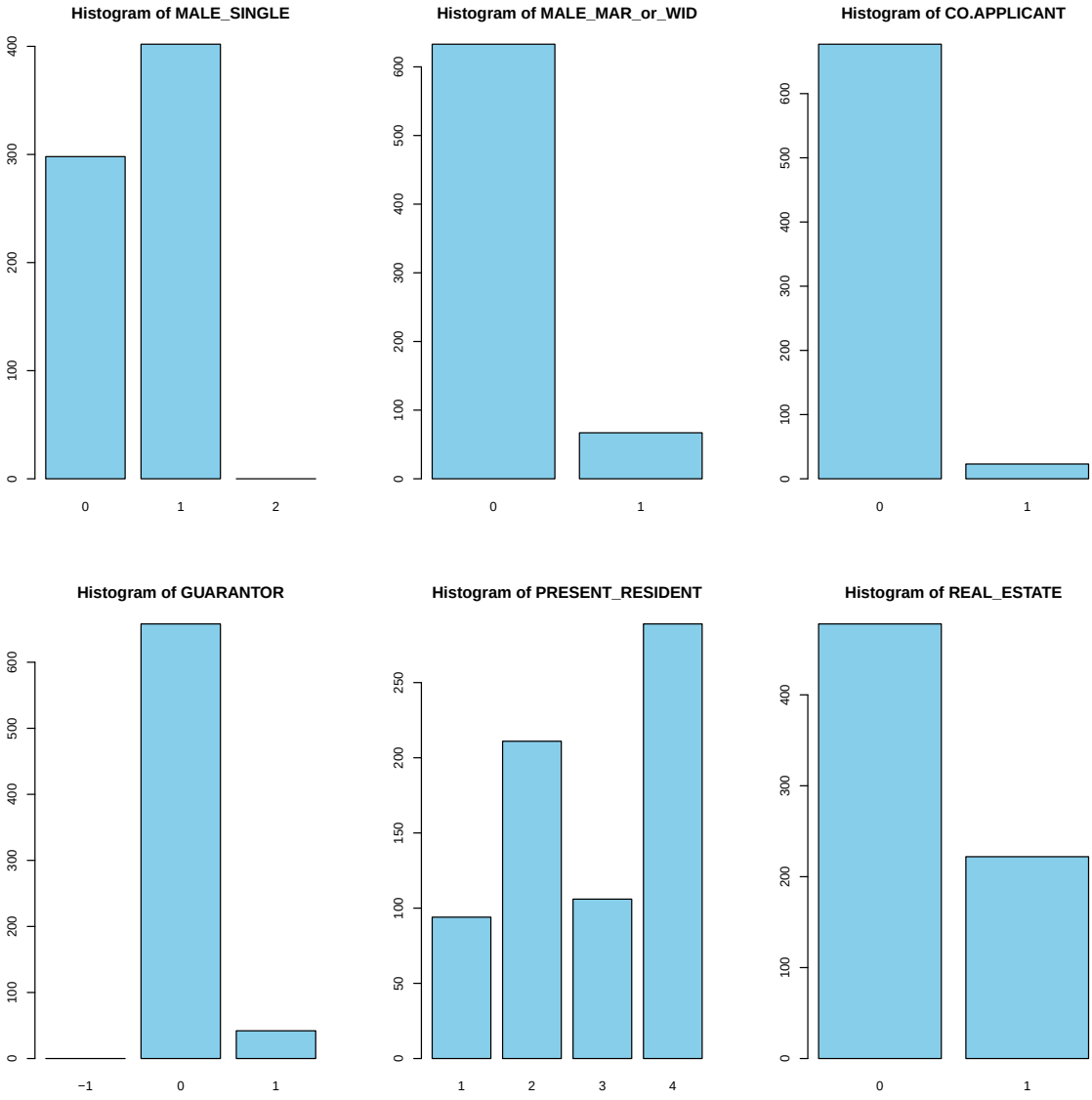

Histogram of HISTORY
0
50
100
150
200
250
300
350
0
1
2
3
4

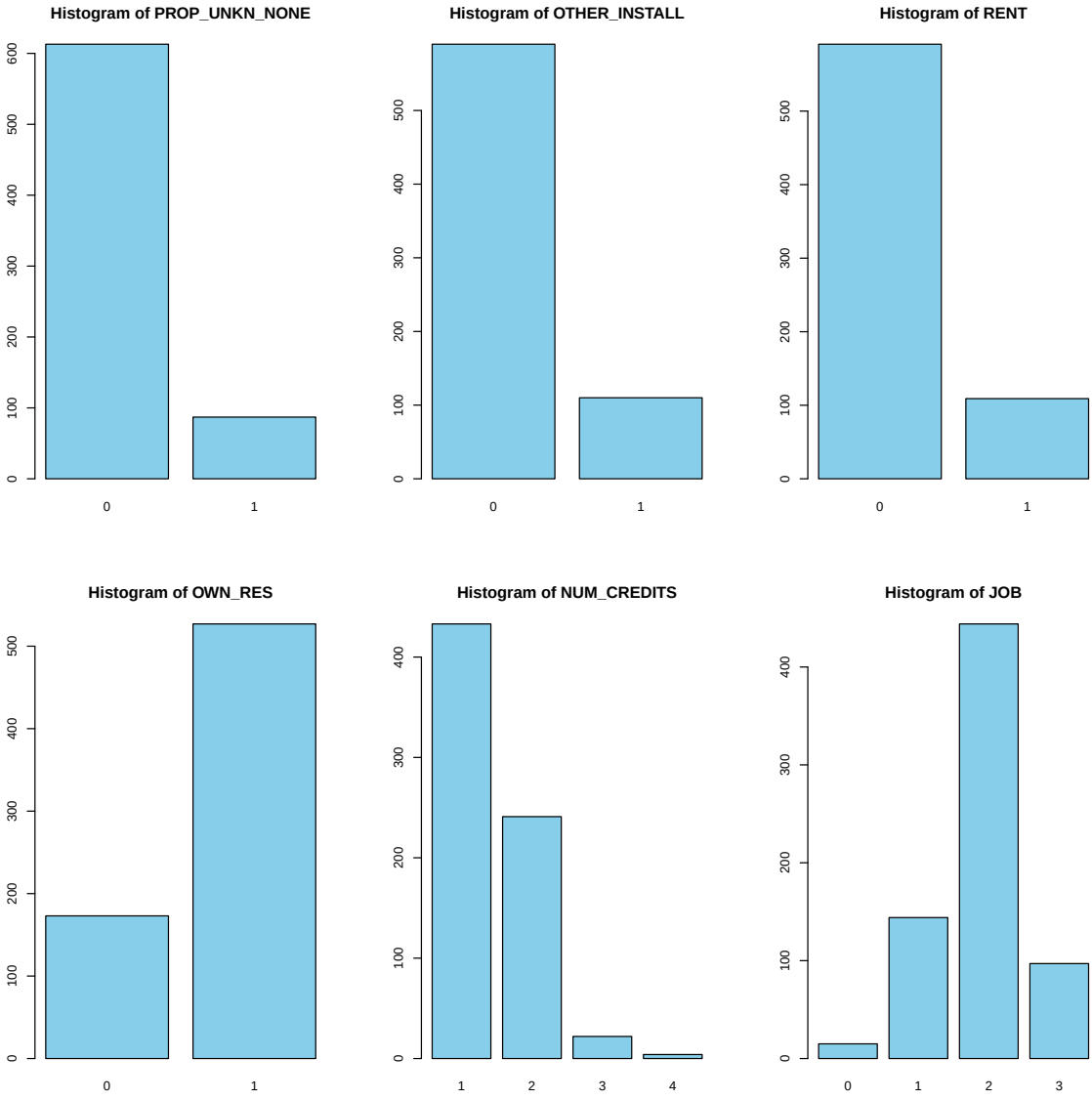

Histogram of NEW_CAR
0
100
200
300
400
500
0
1

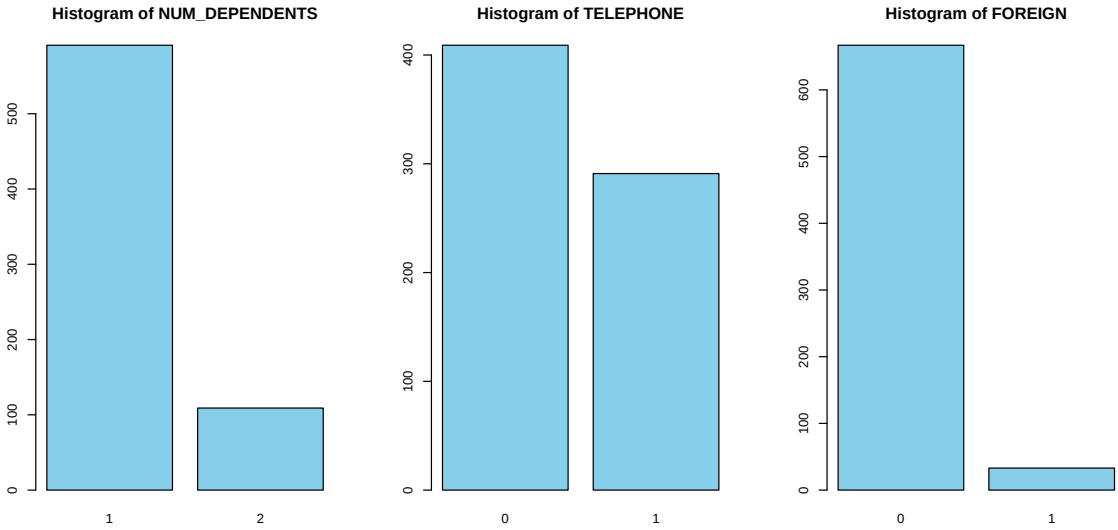

Histogram of USED_CAR
0
100
200
300
400
500
600
0
1

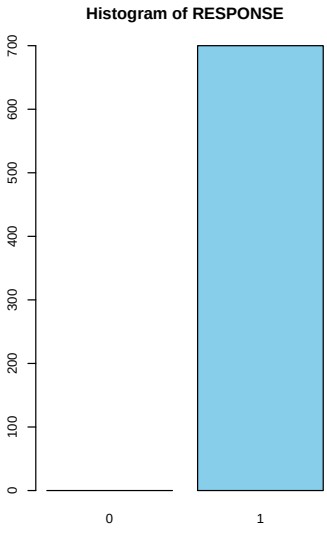

Histogram of FURNITURE
0
100
200
300
400
500
0
1


Histogram of RADIO.TV
0
100
200
300
400
0
1
Histogram of EDUCATION
0
100
200
300
400
500
600
0
1
Histogram of RETRAINING
0
100
200
300
400
500
600
0
1
Histogram of SAV_ACCT
0
100
200
300
0
1
2
3
4
Histogram of EMPLOYMENT
0
50
100
150
200
0
1
2
3
4
Histogram of MALE_DIV
0
100
200
300
400
500
600
0
1


Histogram of MALE_SINGLE
0
100
200
300
400
0
1
2
Histogram of MALE_MAR_or_WID
0
100
200
300
400
500
600
0
1
Histogram of CO.APPLICANT
0
100
200
300
400
500
600
0
1
Histogram of GUARANTOR
0
100
200
300
400
500
600
−1
0
1
Histogram of PRESENT_RESIDENT
0
50
100
150
200
250
1
2
3
4
Histogram of REAL_ESTATE
0
100
200
300
400
0
1

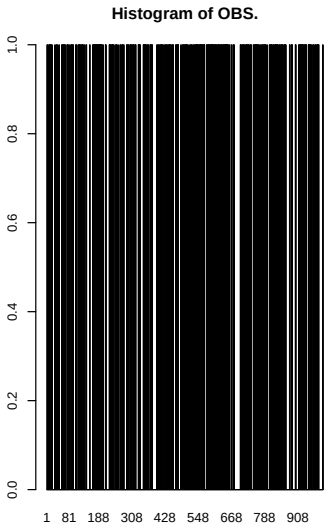

Histogram of PROP_UNKN_NONE
0
100
200
300
400
500
600
0
1
Histogram of OTHER_INSTALL
0
100
200
300
400
500
0
1
Histogram of RENT
0
100
200
300
400
500
0
1
Histogram of OWN_RES
0
100
200
300
400
500
0
1
Histogram of NUM_CREDITS
0
100
200
300
400
1
2
3
4
Histogram of JOB
0
100
200
300
400
0
1
2
3

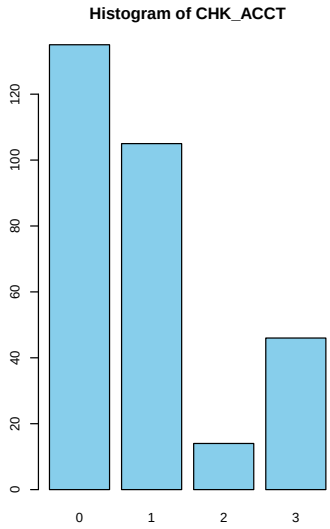


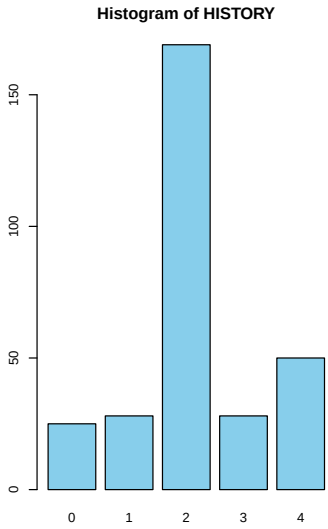


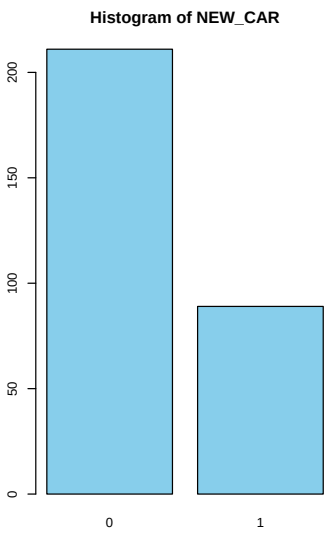

```r
# affichage des histogrammes pour les variables catégorielles
categorical_data_candidat = categorical_data[categorical_data$RESPONSE == 0,]

par(mfrow = c(2, 3))

for (col in names(categorical_data)) {
    barplot(table(categorical_data_candidat[[col]]), main = paste("Histogram of", col), co
```

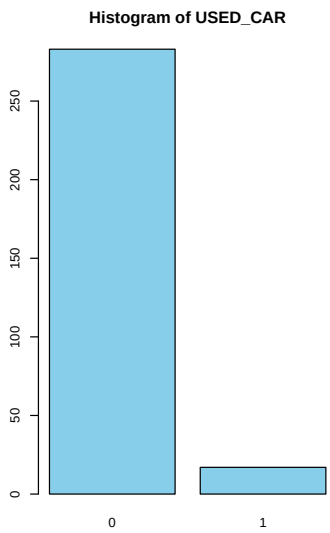
```
}
```

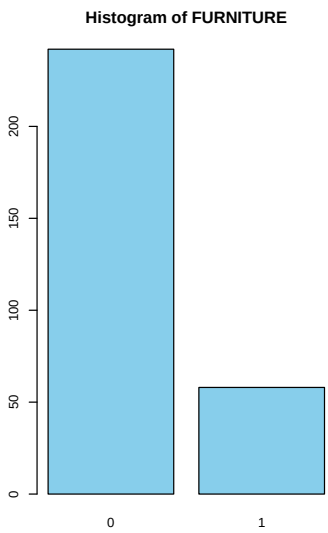

Histogram of OBS.
0.0
0.2
0.4
0.6
0.8
1.0
1
81
188
308
428
548
668
788
908

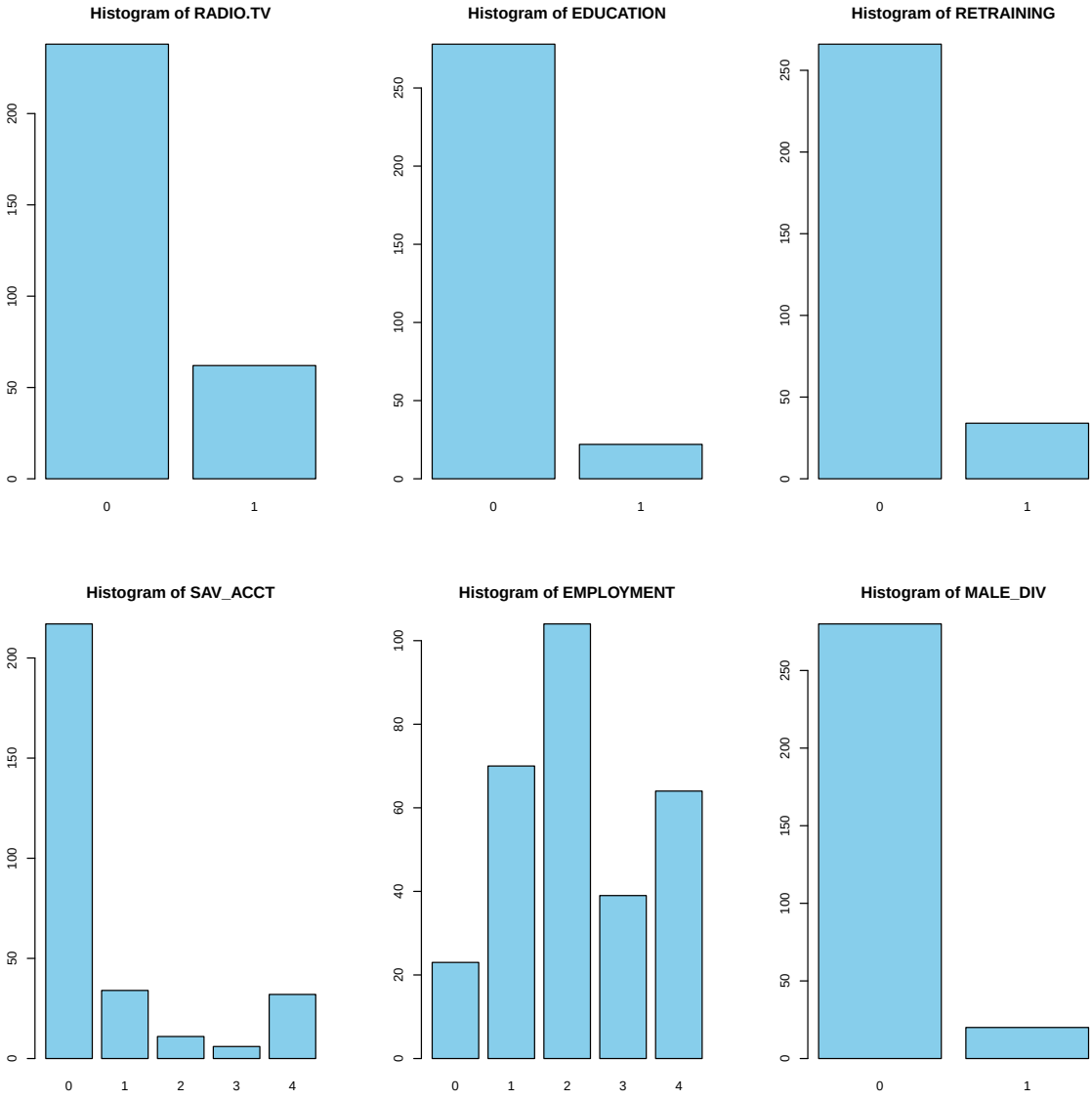

Histogram of CHK_ACCT
0
20
40
60
80
100
120
0
1
2
3

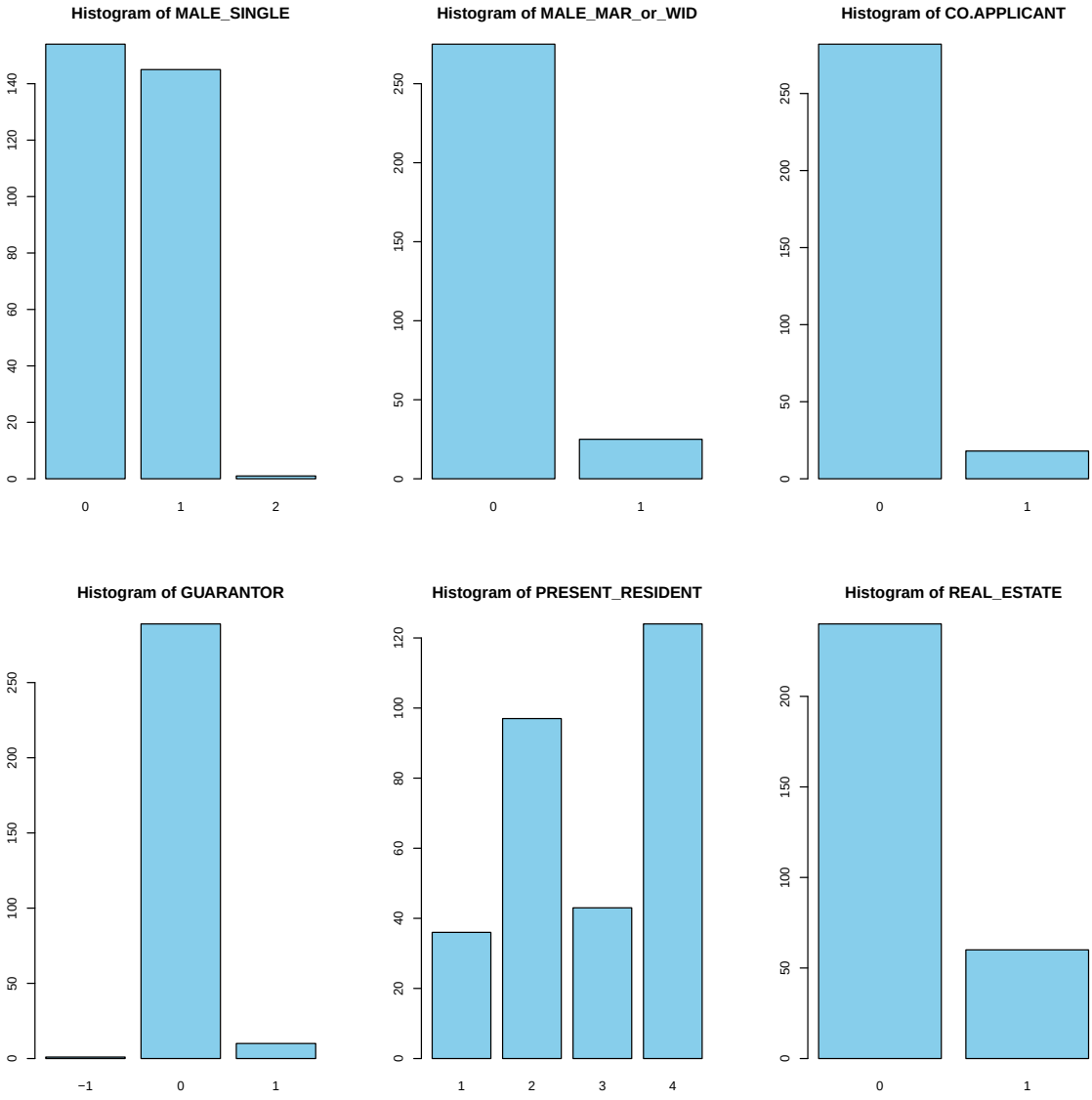

Histogram of HISTORY
0
50
100
150
0
1
2
3
4

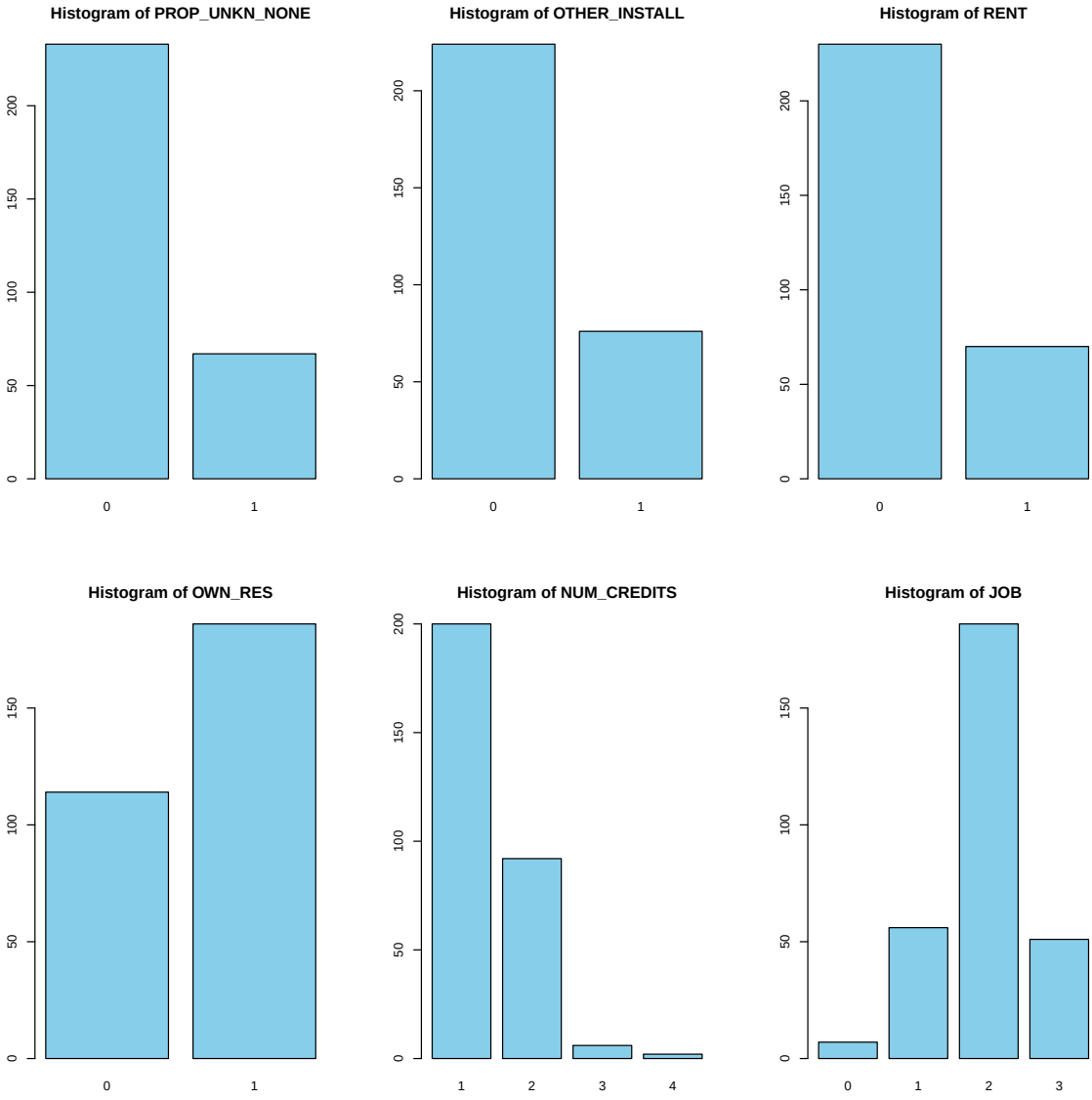

Histogram of NEW_CAR
0
50
100
150
200
0
1

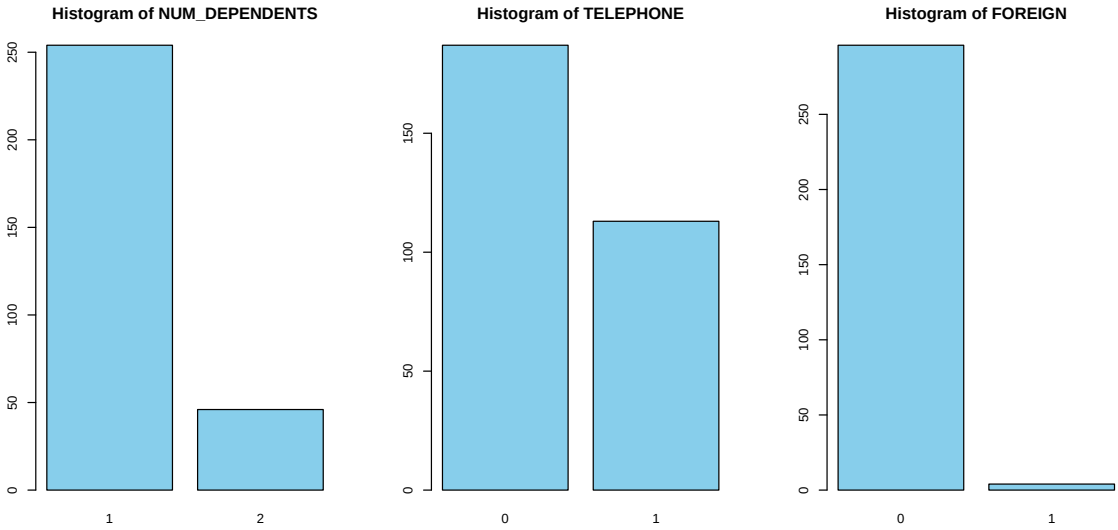

Histogram of USED_CAR
0
50
100
150
200
250
0
1

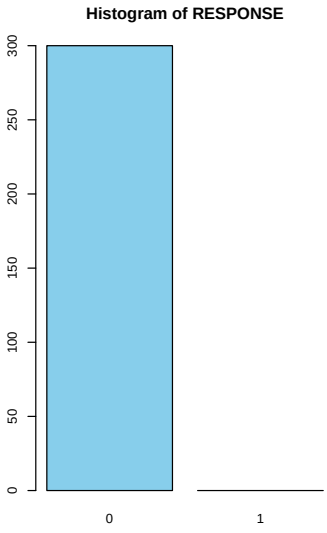

Histogram of FURNITURE
0
50
100
150
200
0
1


Histogram of RADIO.TV
0
50
100
150
200
0
1
Histogram of EDUCATION
0
50
100
150
200
250
0
1
Histogram of RETRAINING
0
50
100
150
200
250
0
1
Histogram of SAV_ACCT
0
50
100
150
200
0
1
2
3
4
Histogram of EMPLOYMENT
0
20
40
60
80
100
0
1
2
3
4
Histogram of MALE_DIV
0
50
100
150
200
250
0
1


Histogram of MALE_SINGLE
0
20
40
60
80
100
120
140
0
1
2
Histogram of MALE_MAR_or_WID
0
50
100
150
200
250
0
1
Histogram of CO.APPLICANT
0
50
100
150
200
250
0
1
Histogram of GUARANTOR
0
50
100
150
200
250
−1
0
1
Histogram of PRESENT_RESIDENT
0
20
40
60
80
100
120
1
2
3
4
Histogram of REAL_ESTATE
0
50
100
150
200
0
1

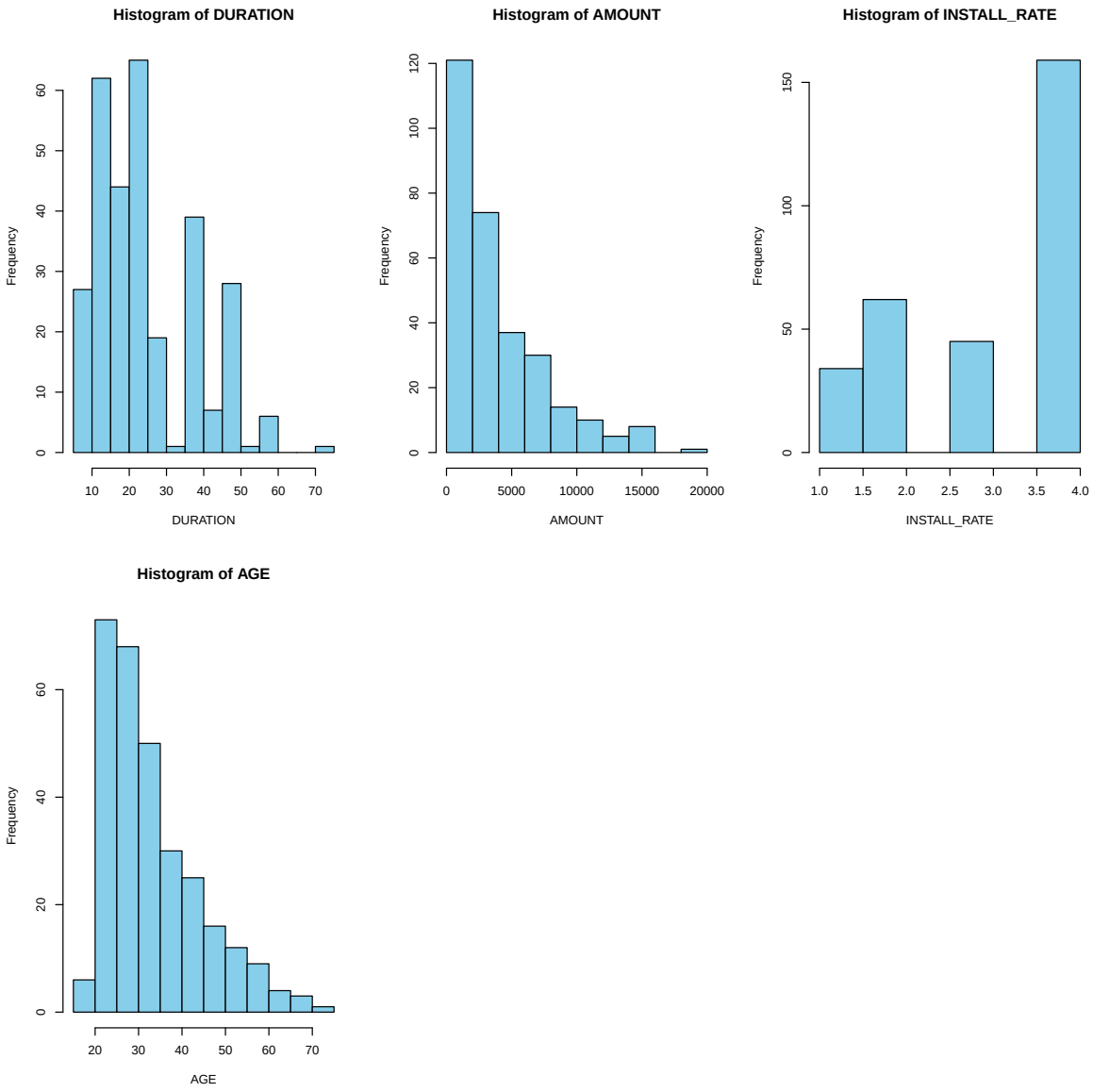

Histogram of PROP_UNKN_NONE
0
50
100
150
200
0
1
Histogram of OTHER_INSTALL
0
50
100
150
200
0
1
Histogram of RENT
0
50
100
150
200
0
1
Histogram of OWN_RES
0
50
100
150
0
1
Histogram of NUM_CREDITS
0
50
100
150
200
1
2
3
4
Histogram of JOB
0
50
100
150
0
1
2
3



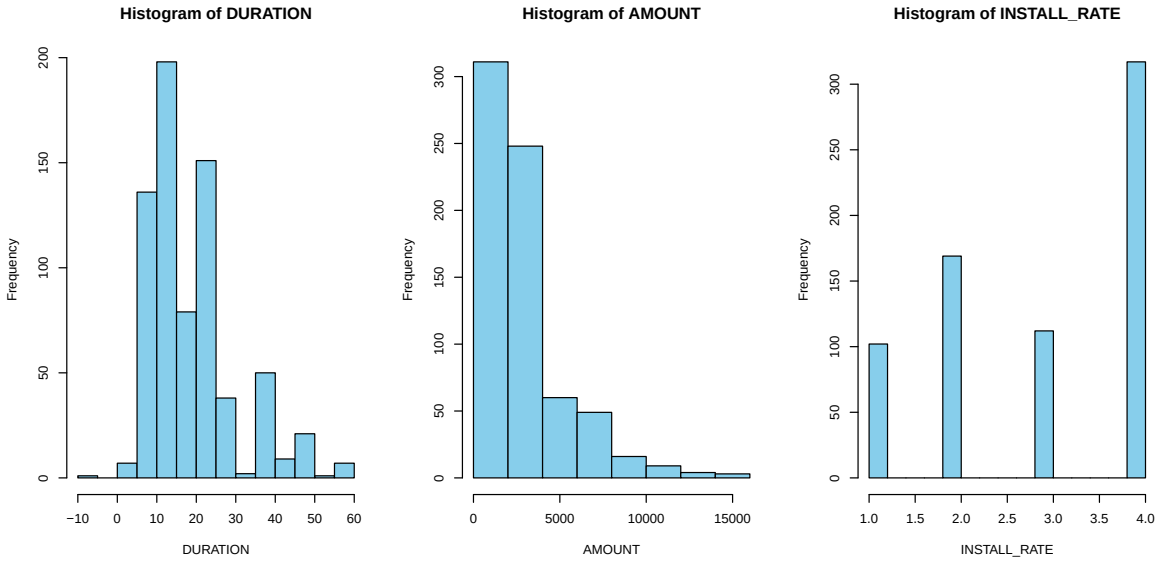


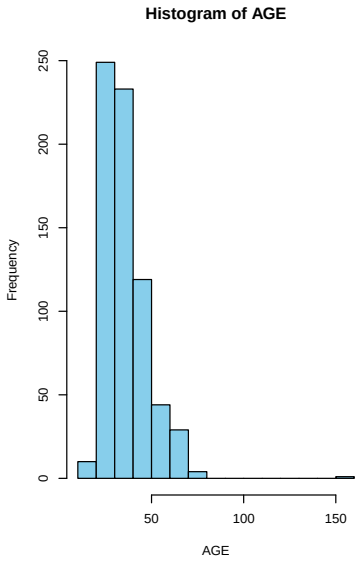

```
numeric_data_candidat = german_credit[german_credit$RESPONSE == 0,]
numeric_data_candidat = numeric_data_candidat[quanti]


par(mfrow = c(2, 3))
```

```r
for (col in names(numeric_data)) {
    hist(numeric_data_candidat[[col]], main = paste("Histogram of", col), xlab = col, col
}
par(mfrow = c(1, 1))
```



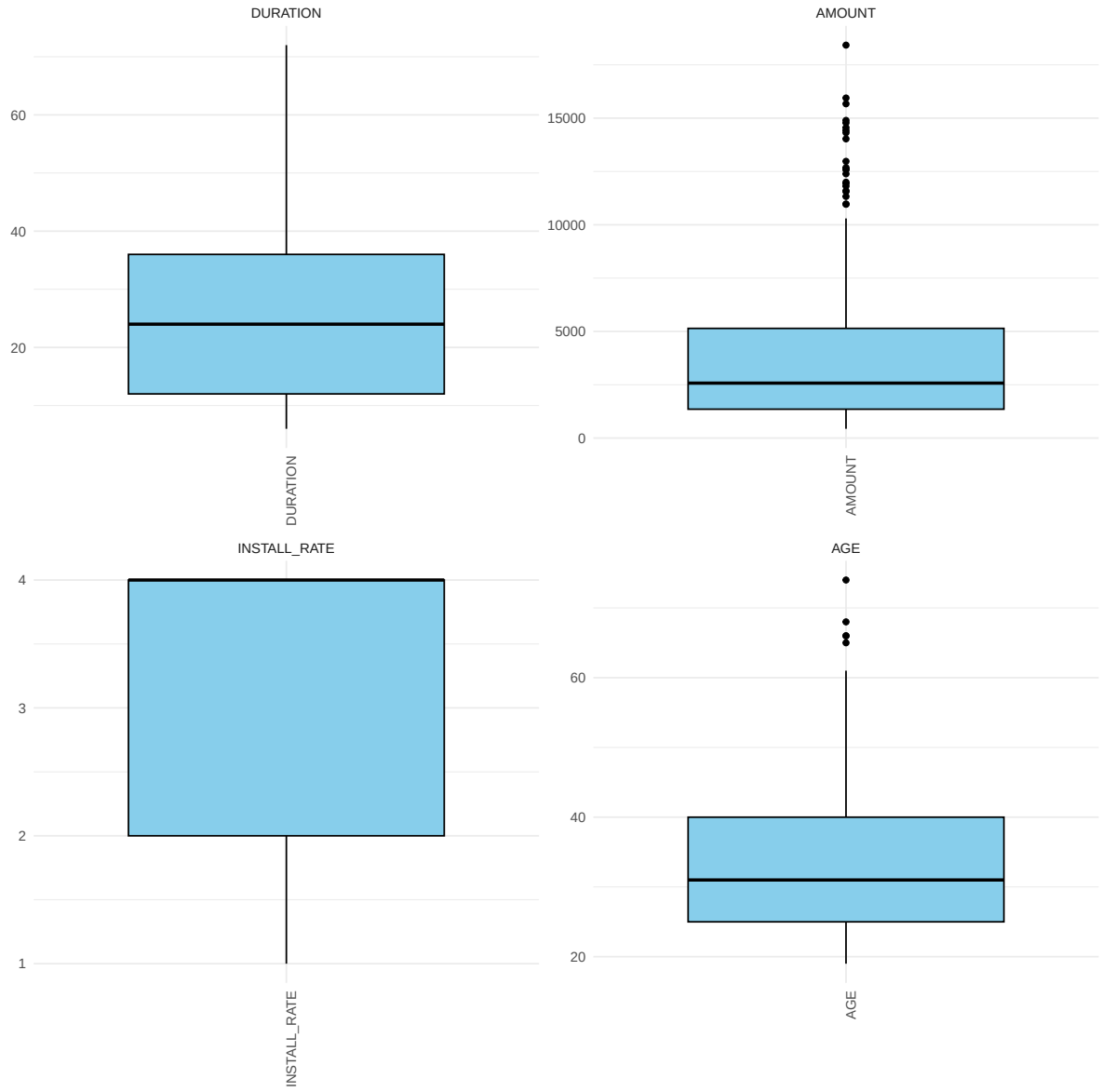
```r
numeric_data_candidat = german_credit[german_credit$RESPONSE == 1,]
numeric_data_candidat = numeric_data_candidat[quanti]



par(mfrow = c(2, 3))

for (col in names(numeric_data)) {
    hist(numeric_data_candidat[[col]], main = paste("Histogram of", col), xlab = col, col
}
par(mfrow = c(1, 1))
```






```
numeric_data_candidat = german_credit[german_credit$RESPONSE == 0,]
numeric_data_candidat = numeric_data_candidat[quanti]

library(ggplot2)
par(mfrow = c(1, 1))

p <- ggplot(stack(numeric_data_candidat), aes(x = ind, y = values)) +
```

```
      geom_boxplot(fill = "skyblue", color = "black") +
      labs(x = NULL, y = NULL) +
      theme_minimal() +
      theme(axis.text.x = element_text(angle = 90, hjust = 1),
            axis.ticks.x = element_blank())

p <- p + facet_wrap(~ind, scales = "free")
print(p)
```



```r
summary(numeric_data$AGE)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  19.00   27.00   33.00   35.53   42.00  151.00      14
```

```r
summary(numeric_data$DURATION)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  -6.00   12.00   18.00   20.89   24.00   72.00
```

Correction des valeurs aberrantes

```r
numeric_data$AGE[numeric_data$AGE > 100 ] <- 75
numeric_data$DURATION[numeric_data$DURATION < 0] <- 0

##passage en format numérique
categorical_data$PRESENT_RESIDENT <- as.numeric(categorical_data$PRESENT_RESIDENT)

##remplacement des valeurs
categorical_data$MALE_SINGLE[categorical_data$MALE_SINGLE == 2] <- 1
categorical_data$GUARANTOR[categorical_data$GUARANTOR == -1] <- 1
categorical_data$PRESENT_RESIDENT <- categorical_data$PRESENT_RESIDENT - 1

#passage en format catégorique
categorical_data$MALE_SINGLE <- as.factor(categorical_data$MALE_SINGLE)
categorical_data$GUARANTOR <- as.factor(categorical_data$GUARANTOR)
categorical_data$PRESENT_RESIDENT <- as.factor(categorical_data$PRESENT_RESIDENT)
```

Imputation des valeurs manquantes

```r
library(zoo)
```

```
Attaching package: 'zoo'

The following objects are masked from 'package:base':

    as.Date, as.Date.numeric
```

```r
missing_values <- is.na(german_credit)
missing_count <- colSums(missing_values)
# imputation des valeurs manquantes avec la médiane
numeric_data <- na.aggregate(numeric_data, FUN = median)
```

TODO: traiter les valeurs aberrantes!

```r
# recherche des meilleures variable pour régression linéaire
# concaténation des deux dataframes

combined_data <- cbind(categorical_data, numeric_data)
# suppression de la colonne OBS
combined_data = combined_data[,-1]
model <- glm(combined_data$RESPONSE ~ ., data=combined_data, family=binomial)

summary(model)
```

```
Call:
glm(formula = combined_data$RESPONSE ~ ., family = binomial,
    data = combined_data)

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        1.508e+00  1.062e+00   1.421 0.155414
CHK_ACCT1          3.829e-01  2.166e-01   1.768 0.077076 .
CHK_ACCT2          9.824e-01  3.707e-01   2.650 0.008045 **
CHK_ACCT3          1.763e+00  2.334e-01   7.553 4.24e-14 ***
HISTORY1          -8.221e-02  5.525e-01  -0.149 0.881713
HISTORY2           5.716e-01  4.393e-01   1.301 0.193208
HISTORY3           9.502e-01  4.746e-01   2.002 0.045298 *
HISTORY4           1.499e+00  4.422e-01   3.390 0.000700 ***
NEW_CAR1          -7.614e-01  3.882e-01  -1.961 0.049861 *
USED_CAR1          8.661e-01  4.882e-01   1.774 0.076029 .
FURNITURE1        -2.158e-02  4.050e-01  -0.053 0.957505
RADIO.TV1          1.310e-01  3.930e-01   0.333 0.738807
EDUCATION1        -8.960e-01  5.059e-01  -1.771 0.076504 .
RETRAINING1       -7.495e-02  4.501e-01  -0.167 0.867755
SAV_ACCT1          3.477e-01  2.906e-01   1.196 0.231546
SAV_ACCT2          3.808e-01  4.006e-01   0.951 0.341746
SAV_ACCT3          1.404e+00  5.374e-01   2.612 0.009005 **
SAV_ACCT4          9.787e-01  2.629e-01   3.723 0.000197 ***
EMPLOYMENT1       -8.985e-02  4.351e-01  -0.207 0.836383
```

```
EMPLOYMENT2        2.306e-01  4.166e-01   0.554 0.579829
EMPLOYMENT3        7.544e-01  4.517e-01   1.670 0.094856 .
EMPLOYMENT4        2.397e-01  4.196e-01   0.571 0.567899
MALE_DIV1         -2.825e-01  3.887e-01  -0.727 0.467404
MALE_SINGLE1       5.705e-01  2.112e-01   2.702 0.006892 **
MALE_MAR_or_WID1   1.496e-01  3.146e-01   0.476 0.634423
CO.APPLICANT1     -3.889e-01  4.074e-01  -0.955 0.339766
GUARANTOR1         8.289e-01  4.061e-01   2.041 0.041212 *
PRESENT_RESIDENT1 -7.654e-01  2.972e-01  -2.575 0.010021 *
PRESENT_RESIDENT2 -4.783e-01  3.321e-01  -1.440 0.149801
PRESENT_RESIDENT3 -3.888e-01  3.009e-01  -1.292 0.196411
REAL_ESTATE1       1.987e-01  2.155e-01   0.922 0.356378
PROP_UNKN_NONE1   -5.647e-01  3.899e-01  -1.448 0.147560
OTHER_INSTALL1    -5.884e-01  2.140e-01  -2.750 0.005958 **
RENT1             -6.308e-01  4.823e-01  -1.308 0.190888
OWN_RES1          -1.819e-01  4.579e-01  -0.397 0.691237
NUM_CREDITS2      -3.975e-01  2.438e-01  -1.630 0.103021
NUM_CREDITS3      -3.257e-01  6.050e-01  -0.538 0.590334
NUM_CREDITS4      -5.310e-01  1.098e+00  -0.484 0.628708
JOB1              -3.847e-01  6.757e-01  -0.569 0.569151
JOB2              -4.250e-01  6.532e-01  -0.651 0.515335
JOB3              -2.720e-01  6.601e-01  -0.412 0.680231
NUM_DEPENDENTS2   -2.544e-01  2.504e-01  -1.016 0.309691
TELEPHONE1         2.919e-01  2.010e-01   1.452 0.146489
FOREIGN1           1.465e+00  6.285e-01   2.331 0.019774 *
DURATION          -2.808e-02  9.400e-03  -2.987 0.002813 **
AMOUNT            -1.160e-04  4.489e-05  -2.585 0.009733 **
INSTALL_RATE      -3.210e-01  8.878e-02  -3.615 0.000300 ***
AGE                1.260e-02  9.447e-03   1.334 0.182318
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1221.73  on 999  degrees of freedom
Residual deviance:  892.11  on 952  degrees of freedom
AIC: 988.11

Number of Fisher Scoring iterations: 5
```