Brouillon-project-in-ADA

Elodie Kwan

2022-04-14

Import libraries and data

Download libraries and set the working directory

```
# library(here)
library(dplyr)
```

```
##
## Attachement du package : 'dplyr'

## Les objets suivants sont masqués depuis 'package:stats':
##
##     filter, lag

## Les objets suivants sont masqués depuis 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(Hmisc)
```

```
## Le chargement a nécessité le package : lattice

## Le chargement a nécessité le package : survival

## Le chargement a nécessité le package : Formula

## Le chargement a nécessité le package : ggplot2

##
## Attachement du package : 'Hmisc'

## Les objets suivants sont masqués depuis 'package:dplyr':
##
##     src, summarize

## Les objets suivants sont masqués depuis 'package:base':
##
##     format.pval, units
```

```
library(DataExplorer)
library(psych)
```

```
##
## Attachement du package : 'psych'

## L'objet suivant est masqué depuis 'package:Hmisc':
##
##     describe
```

```
## Les objets suivants sont masqués depuis 'package:ggplot2':
##
##     %+%, alpha
```

```r
library(rpart)
```

Import data

```r
German_credit <- read.csv("./Data_DA/GermanCredit.csv", header = TRUE, sep = ";")
```

Get to know the data

Title : german credit data

Name of the file : GermanCredit.cvs

Abstract

The German Credit data has data on 1000 past credit applicants, described by 30 variables. Each applicant is rated as “Good” or “Bad” credit (encoded as 1 and 0 respectively in the response variable). **Goal** : We want to obtain a model that may be used to determine if new applicants present a good or bad credit risk

- Number of instances : 1000
- Number of attributes : 30
- Attribute Information :

```r
head(German_credit, 10)
```

```
##    OBS. CHK_ACCT DURATION HISTORY NEW_CAR USED_CAR FURNITURE RADIO.TV EDUCATION
## 1     1        0        6       4       0        0         0        1         0
## 2     2        1       48       2       0        0         0        1         0
## 3     3        3       12       4       0        0         0        0         1
## 4     4        0       42       2       0        0         1        0         0
## 5     5        0       24       3       1        0         0        0         0
## 6     6        3       36       2       0        0         0        0         1
## 7     7        3       24       2       0        0         1        0         0
## 8     8        1       36       2       0        1         0        0         0
## 9     9        3       12       2       0        0         0        1         0
## 10   10        1       30       4       1        0         0        0         0
##    RETRAINING AMOUNT SAV_ACCT EMPLOYMENT INSTALL_RATE MALE_DIV MALE_SINGLE
## 1           0   1169        4          4            4        0           1
## 2           0   5951        0          2            2        0           0
## 3           0   2096        0          3            2        0           1
## 4           0   7882        0          3            2        0           1
## 5           0   4870        0          2            3        0           1
## 6           0   9055        4          2            2        0           1
## 7           0   2835        2          4            3        0           1
## 8           0   6948        0          2            2        0           1
## 9           0   3059        3          3            2        1           0
## 10          0   5234        0          0            4        0           0
##    MALE_MAR_or_WID CO.APPLICANT GUARANTOR PRESENT_RESIDENT REAL_ESTATE
## 1                0            0         0                4           1
## 2                0            0         0                2           1
## 3                0            0         0                3           1
## 4                0            0         1                4           0
## 5                0            0         0                4           0
## 6                0            0         0                4           0
## 7                0            0         0                4           0
```

```
## 8                0             0         0                2           0
## 9                0             0         0                4           1
## 10               1             0         0                2           0
##    PROP_UNKN_NONE AGE OTHER_INSTALL RENT OWN_RES NUM_CREDITS JOB NUM_DEPENDENTS
## 1               0  67             0    0       1           2   2              1
## 2               0  22             0    0       1           1   2              1
## 3               0  49             0    0       1           1   1              2
## 4               0  45             0    0       0           1   2              2
## 5               1  53             0    0       0           2   2              2
## 6               1  35             0    0       0           1   1              2
## 7               0  53             0    0       1           1   2              1
## 8               0  35             0    1       0           1   3              1
## 9               0  61             0    0       1           1   1              1
## 10              0  28             0    0       1           2   3              1
##    TELEPHONE FOREIGN RESPONSE
## 1          1       0        1
## 2          0       0        0
## 3          0       0        1
## 4          0       0        1
## 5          0       0        0
## 6          1       0        1
## 7          0       0        1
## 8          1       0        1
## 9          0       0        1
## 10         0       0        0
```

```
summary(German_credit)
```

```
##       OBS.           CHK_ACCT        DURATION       HISTORY
##  Min.   :   1.0   Min.   :0.000   Min.   : 4.0   Min.   :0.000
##  1st Qu.: 250.8   1st Qu.:0.000   1st Qu.:12.0   1st Qu.:2.000
##  Median : 500.5   Median :1.000   Median :18.0   Median :2.000
##  Mean   : 500.5   Mean   :1.577   Mean   :20.9   Mean   :2.545
##  3rd Qu.: 750.2   3rd Qu.:3.000   3rd Qu.:24.0   3rd Qu.:4.000
##  Max.   :1000.0   Max.   :3.000   Max.   :72.0   Max.   :4.000
##     NEW_CAR         USED_CAR       FURNITURE        RADIO.TV
##  Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.00
##  1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.00
##  Median :0.000   Median :0.000   Median :0.000   Median :0.00
##  Mean   :0.234   Mean   :0.103   Mean   :0.181   Mean   :0.28
##  3rd Qu.:0.000   3rd Qu.:0.000   3rd Qu.:0.000   3rd Qu.:1.00
##  Max.   :1.000   Max.   :1.000   Max.   :1.000   Max.   :1.00
##    EDUCATION        RETRAINING        AMOUNT        SAV_ACCT
##  Min.   :-1.000   Min.   :0.000   Min.   :  250   Min.   :0.000
##  1st Qu.: 0.000   1st Qu.:0.000   1st Qu.: 1366   1st Qu.:0.000
##  Median : 0.000   Median :0.000   Median : 2320   Median :0.000
##  Mean   : 0.048   Mean   :0.097   Mean   : 3271   Mean   :1.105
##  3rd Qu.: 0.000   3rd Qu.:0.000   3rd Qu.: 3972   3rd Qu.:2.000
##  Max.   : 1.000   Max.   :1.000   Max.   :18424   Max.   :4.000
##    EMPLOYMENT     INSTALL_RATE      MALE_DIV     MALE_SINGLE    MALE_MAR_or_WID
##  Min.   :0.000   Min.   :1.000   Min.   :0.00   Min.   :0.000   Min.   :0.000
##  1st Qu.:2.000   1st Qu.:2.000   1st Qu.:0.00   1st Qu.:0.000   1st Qu.:0.000
##  Median :2.000   Median :3.000   Median :0.00   Median :1.000   Median :0.000
##  Mean   :2.384   Mean   :2.973   Mean   :0.05   Mean   :0.548   Mean   :0.092
##  3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:0.00   3rd Qu.:1.000   3rd Qu.:0.000
```

```
##  Max.   :4.000   Max.   :4.000   Max.   :1.00   Max.   :1.000   Max.   :1.000
##   CO.APPLICANT     GUARANTOR     PRESENT_RESIDENT  REAL_ESTATE
##  Min.   :0.000   Min.   :0.000   Min.   :1.000    Min.   :0.000
##  1st Qu.:0.000   1st Qu.:0.000   1st Qu.:2.000    1st Qu.:0.000
##  Median :0.000   Median :0.000   Median :3.000    Median :0.000
##  Mean   :0.041   Mean   :0.053   Mean   :2.845    Mean   :0.282
##  3rd Qu.:0.000   3rd Qu.:0.000   3rd Qu.:4.000    3rd Qu.:1.000
##  Max.   :1.000   Max.   :2.000   Max.   :4.000    Max.   :1.000
##  PROP_UNKN_NONE       AGE        OTHER_INSTALL        RENT
##  Min.   :0.000   Min.   : 19.0   Min.   :0.000   Min.   :0.000
##  1st Qu.:0.000   1st Qu.: 27.0   1st Qu.:0.000   1st Qu.:0.000
##  Median :0.000   Median : 33.0   Median :0.000   Median :0.000
##  Mean   :0.154   Mean   : 35.6   Mean   :0.186   Mean   :0.179
##  3rd Qu.:0.000   3rd Qu.: 42.0   3rd Qu.:0.000   3rd Qu.:0.000
##  Max.   :1.000   Max.   :125.0   Max.   :1.000   Max.   :1.000
##     OWN_RES       NUM_CREDITS         JOB        NUM_DEPENDENTS
##  Min.   :0.000   Min.   :1.000   Min.   :0.000   Min.   :1.000
##  1st Qu.:0.000   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:1.000
##  Median :1.000   Median :1.000   Median :2.000   Median :1.000
##  Mean   :0.713   Mean   :1.407   Mean   :1.904   Mean   :1.155
##  3rd Qu.:1.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:1.000
##  Max.   :1.000   Max.   :4.000   Max.   :3.000   Max.   :2.000
##    TELEPHONE        FOREIGN         RESPONSE
##  Min.   :0.000   Min.   :0.000   Min.   :0.0
##  1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.0
##  Median :0.000   Median :0.000   Median :1.0
##  Mean   :0.404   Mean   :0.037   Mean   :0.7
##  3rd Qu.:1.000   3rd Qu.:0.000   3rd Qu.:1.0
##  Max.   :1.000   Max.   :1.000   Max.   :1.0
```

```
str(German_credit)
```

```
## 'data.frame':    1000 obs. of  32 variables:
##  $ OBS.            : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ CHK_ACCT        : int  0 1 3 0 0 3 3 1 3 1 ...
##  $ DURATION        : int  6 48 12 42 24 36 24 36 12 30 ...
##  $ HISTORY         : int  4 2 4 2 3 2 2 2 2 4 ...
##  $ NEW_CAR         : int  0 0 0 0 1 0 0 0 0 1 ...
##  $ USED_CAR        : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ FURNITURE       : int  0 0 0 1 0 0 1 0 0 0 ...
##  $ RADIO.TV        : int  1 1 0 0 0 0 0 0 1 0 ...
##  $ EDUCATION       : int  0 0 1 0 0 1 0 0 0 0 ...
##  $ RETRAINING      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AMOUNT          : int  1169 5951 2096 7882 4870 9055 2835 6948 3059 5234 ...
##  $ SAV_ACCT        : int  4 0 0 0 0 4 2 0 3 0 ...
##  $ EMPLOYMENT      : int  4 2 3 3 2 2 4 2 3 0 ...
##  $ INSTALL_RATE    : int  4 2 2 2 3 2 3 2 2 4 ...
##  $ MALE_DIV        : int  0 0 0 0 0 0 0 0 1 0 ...
##  $ MALE_SINGLE     : int  1 0 1 1 1 1 1 1 0 0 ...
##  $ MALE_MAR_or_WID : int  0 0 0 0 0 0 0 0 0 1 ...
##  $ CO.APPLICANT    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ GUARANTOR       : int  0 0 0 1 0 0 0 0 0 0 ...
##  $ PRESENT_RESIDENT: int  4 2 3 4 4 4 4 2 4 2 ...
##  $ REAL_ESTATE     : int  1 1 1 0 0 0 0 0 1 0 ...
##  $ PROP_UNKN_NONE  : int  0 0 0 0 1 1 0 0 0 0 ...
```

```
##  $ AGE            : int  67 22 49 45 53 35 53 35 61 28 ...
##  $ OTHER_INSTALL  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ RENT           : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ OWN_RES        : int  1 1 1 0 0 0 1 0 1 1 ...
##  $ NUM_CREDITS    : int  2 1 1 1 2 1 1 1 1 2 ...
##  $ JOB            : int  2 2 1 2 2 1 2 3 1 3 ...
##  $ NUM_DEPENDENTS : int  1 1 2 2 2 2 1 1 1 1 ...
##  $ TELEPHONE      : int  1 0 0 0 0 1 0 1 0 0 ...
##  $ FOREIGN        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ RESPONSE       : int  1 0 1 1 0 1 1 1 1 0 ...
```

```r
which(is.na(German_credit))
```

```
## integer(0)
```

- There are no missing values.
- The response variable is the '**Response**' variable - last column on the data.

Response variable : credit rating is good 1. 0 : No 2. 1 : Yes

Make sure that the class of the variables are correct :

```r
German_credit$DURATION <- as.numeric(German_credit$DURATION)
German_credit$AMOUNT <- as.numeric(German_credit$AMOUNT)
German_credit$INSTALL_RATE <- as.numeric(German_credit$INSTALL_RATE)
German_credit$AGE <- as.numeric(German_credit$AGE)
German_credit$NUM_CREDITS <- as.numeric(German_credit$NUM_CREDITS)
German_credit$NUM_DEPENDENTS <- as.numeric(German_credit$NUM_DEPENDENTS)

for (i in 1:ncol(German_credit)){
  if (class(German_credit[,i])=="integer"){
    German_credit[,i] <- factor(German_credit[,i])
    }
}

str(German_credit)
```

```
## 'data.frame':    1000 obs. of  32 variables:
##  $ OBS.            : Factor w/ 1000 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ CHK_ACCT        : Factor w/ 4 levels "0","1","2","3": 1 2 4 1 1 4 4 2 4 2 ...
##  $ DURATION        : num  6 48 12 42 24 36 24 36 12 30 ...
##  $ HISTORY         : Factor w/ 5 levels "0","1","2","3",..: 5 3 5 3 4 3 3 3 3 5 ...
##  $ NEW_CAR         : Factor w/ 2 levels "0","1": 1 1 1 1 2 1 1 1 1 2 ...
##  $ USED_CAR        : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 2 1 1 ...
##  $ FURNITURE       : Factor w/ 2 levels "0","1": 1 1 1 2 1 1 2 1 1 1 ...
##  $ RADIO.TV        : Factor w/ 2 levels "0","1": 2 2 1 1 1 1 1 1 2 1 ...
##  $ EDUCATION       : Factor w/ 3 levels "-1","0","1": 2 2 3 2 2 3 2 2 2 2 ...
##  $ RETRAINING      : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ AMOUNT          : num  1169 5951 2096 7882 4870 ...
##  $ SAV_ACCT        : Factor w/ 5 levels "0","1","2","3",..: 5 1 1 1 1 5 3 1 4 1 ...
##  $ EMPLOYMENT      : Factor w/ 5 levels "0","1","2","3",..: 5 3 4 4 3 3 5 3 4 1 ...
##  $ INSTALL_RATE    : num  4 2 2 2 3 2 3 2 2 4 ...
##  $ MALE_DIV        : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 2 1 ...
##  $ MALE_SINGLE     : Factor w/ 2 levels "0","1": 2 1 2 2 2 2 2 2 1 1 ...
##  $ MALE_MAR_or_WID : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 2 ...
##  $ CO.APPLICANT    : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ GUARANTOR       : Factor w/ 3 levels "0","1","2": 1 1 1 2 1 1 1 1 1 1 ...
```

```
##  $ PRESENT_RESIDENT: Factor w/ 4 levels "1","2","3","4": 4 2 3 4 4 4 4 2 4 2 ...
##  $ REAL_ESTATE     : Factor w/ 2 levels "0","1": 2 2 2 1 1 1 1 1 2 1 ...
##  $ PROP_UNKN_NONE  : Factor w/ 2 levels "0","1": 1 1 1 1 2 2 1 1 1 1 ...
##  $ AGE             : num  67 22 49 45 53 35 53 35 61 28 ...
##  $ OTHER_INSTALL   : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ RENT            : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 2 1 1 ...
##  $ OWN_RES         : Factor w/ 2 levels "0","1": 2 2 2 1 1 1 2 1 2 2 ...
##  $ NUM_CREDITS     : num  2 1 1 1 2 1 1 1 1 2 ...
##  $ JOB             : Factor w/ 4 levels "0","1","2","3": 3 3 2 3 3 2 3 4 2 4 ...
##  $ NUM_DEPENDENTS  : num  1 1 2 2 2 2 1 1 1 1 ...
##  $ TELEPHONE       : Factor w/ 2 levels "0","1": 2 1 1 1 1 2 1 2 1 1 ...
##  $ FOREIGN         : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ RESPONSE        : Factor w/ 2 levels "0","1": 2 1 2 2 1 2 2 2 2 1 ...
```

```
describe(German_credit)
```

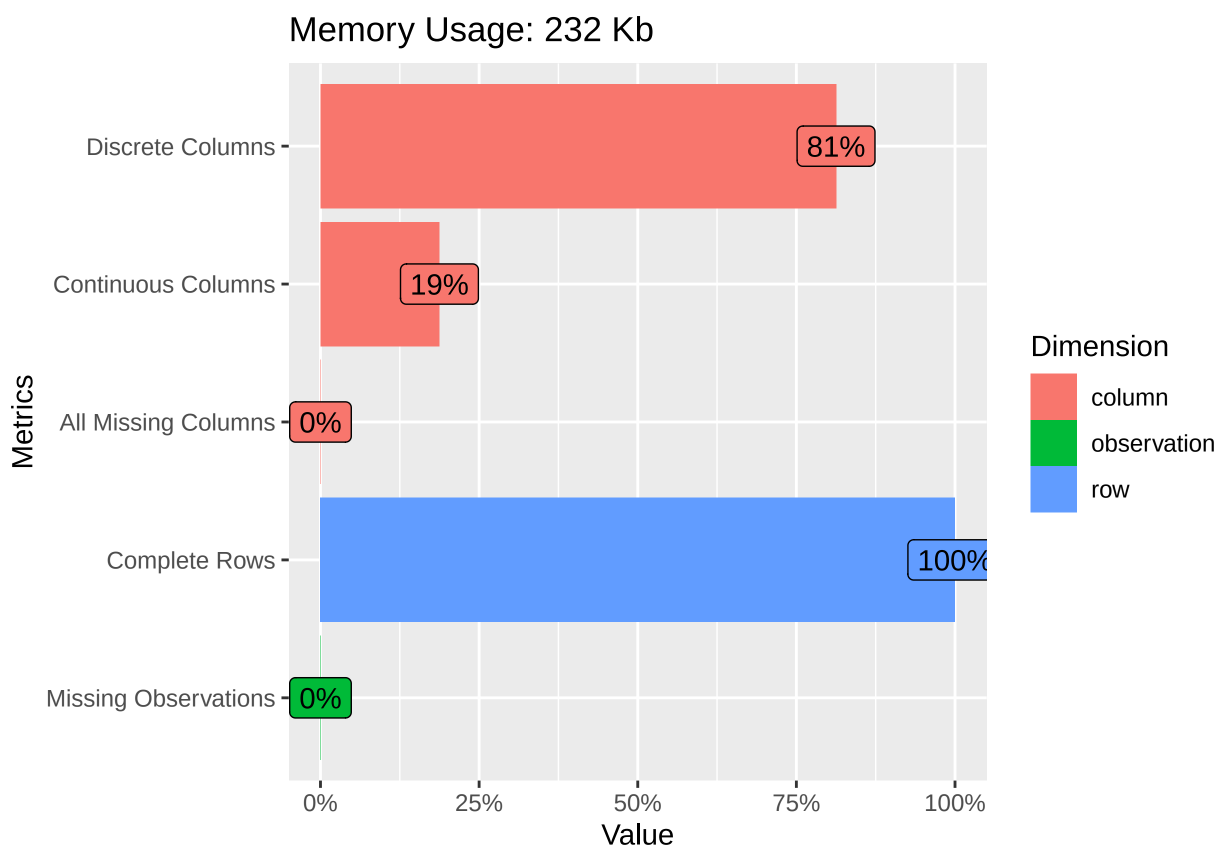
```
##                   vars    n    mean      sd median trimmed     mad min   max
## OBS.*                1 1000  500.50  288.82  500.5  500.50  370.65   1  1000
## CHK_ACCT*            2 1000    2.58    1.26    2.0    2.60    1.48   1     4
## DURATION             3 1000   20.90   12.06   18.0   19.47    8.90   4    72
## HISTORY*             4 1000    3.54    1.08    3.0    3.59    0.00   1     5
## NEW_CAR*             5 1000    1.23    0.42    1.0    1.17    0.00   1     2
## USED_CAR*            6 1000    1.10    0.30    1.0    1.00    0.00   1     2
## FURNITURE*           7 1000    1.18    0.39    1.0    1.10    0.00   1     2
## RADIO.TV*            8 1000    1.28    0.45    1.0    1.23    0.00   1     2
## EDUCATION*           9 1000    2.05    0.22    2.0    2.00    0.00   1     3
## RETRAINING*         10 1000    1.10    0.30    1.0    1.00    0.00   1     2
## AMOUNT              11 1000 3271.26 2822.74 2319.5 2754.57 1627.15 250 18424
## SAV_ACCT*           12 1000    2.10    1.58    1.0    1.88    0.00   1     5
## EMPLOYMENT*         13 1000    3.38    1.21    3.0    3.43    1.48   1     5
## INSTALL_RATE        14 1000    2.97    1.12    3.0    3.09    1.48   1     4
## MALE_DIV*           15 1000    1.05    0.22    1.0    1.00    0.00   1     2
## MALE_SINGLE*        16 1000    1.55    0.50    2.0    1.56    0.00   1     2
## MALE_MAR_or_WID*    17 1000    1.09    0.29    1.0    1.00    0.00   1     2
## CO.APPLICANT*       18 1000    1.04    0.20    1.0    1.00    0.00   1     2
## GUARANTOR*          19 1000    1.05    0.23    1.0    1.00    0.00   1     3
## PRESENT_RESIDENT*   20 1000    2.85    1.10    3.0    2.93    1.48   1     4
## REAL_ESTATE*        21 1000    1.28    0.45    1.0    1.23    0.00   1     2
## PROP_UNKN_NONE*     22 1000    1.15    0.36    1.0    1.07    0.00   1     2
## AGE                 23 1000   35.60   11.66   33.0   34.17   10.38  19   125
## OTHER_INSTALL*      24 1000    1.19    0.39    1.0    1.11    0.00   1     2
## RENT*               25 1000    1.18    0.38    1.0    1.10    0.00   1     2
## OWN_RES*            26 1000    1.71    0.45    2.0    1.77    0.00   1     2
## NUM_CREDITS         27 1000    1.41    0.58    1.0    1.33    0.00   1     4
## JOB*                28 1000    2.90    0.65    3.0    2.91    0.00   1     4
## NUM_DEPENDENTS      29 1000    1.16    0.36    1.0    1.07    0.00   1     2
## TELEPHONE*          30 1000    1.40    0.49    1.0    1.38    0.00   1     2
## FOREIGN*            31 1000    1.04    0.19    1.0    1.00    0.00   1     2
## RESPONSE*           32 1000    1.70    0.46    2.0    1.75    0.00   1     2
##                   range  skew kurtosis   se
## OBS.*               999  0.00    -1.20 9.13
## CHK_ACCT*             3  0.01    -1.66 0.04
## DURATION             68  1.09     0.90 0.38
## HISTORY*              4 -0.01    -0.59 0.03
## NEW_CAR*              1  1.25    -0.43 0.01
```

```
## USED_CAR*              1  2.61     4.81  0.01
## FURNITURE*             1  1.65     0.74  0.01
## RADIO.TV*              1  0.98    -1.04  0.01
## EDUCATION*             2  3.93    15.19  0.01
## RETRAINING*            1  2.72     5.40  0.01
## AMOUNT             18174  1.94     4.25 89.26
## SAV_ACCT*              4  1.01    -0.69  0.05
## EMPLOYMENT*            4 -0.12    -0.94  0.04
## INSTALL_RATE           3 -0.53    -1.21  0.04
## MALE_DIV*              1  4.12    15.02  0.01
## MALE_SINGLE*           1 -0.19    -1.96  0.02
## MALE_MAR_or_WID*       1  2.82     5.95  0.01
## CO.APPLICANT*          1  4.62    19.39  0.01
## GUARANTOR*             2  4.23    17.30  0.01
## PRESENT_RESIDENT*      3 -0.27    -1.38  0.03
## REAL_ESTATE*           1  0.97    -1.07  0.01
## PROP_UNKN_NONE*        1  1.91     1.67  0.01
## AGE                  106  1.35     3.56  0.37
## OTHER_INSTALL*         1  1.61     0.60  0.01
## RENT*                  1  1.67     0.80  0.01
## OWN_RES*               1 -0.94    -1.12  0.01
## NUM_CREDITS            3  1.27     1.58  0.02
## JOB*                   3 -0.37     0.49  0.02
## NUM_DEPENDENTS         1  1.90     1.63  0.01
## TELEPHONE*             1  0.39    -1.85  0.02
## FOREIGN*               1  4.90    22.02  0.01
## RESPONSE*              1 -0.87    -1.24  0.01
```

```
introduce(German_credit)
```

```
##   rows columns discrete_columns continuous_columns all_missing_columns
## 1 1000      32               26                  6                   0
##   total_missing_values complete_rows total_observations memory_usage
## 1                    0          1000              32000       237568
```

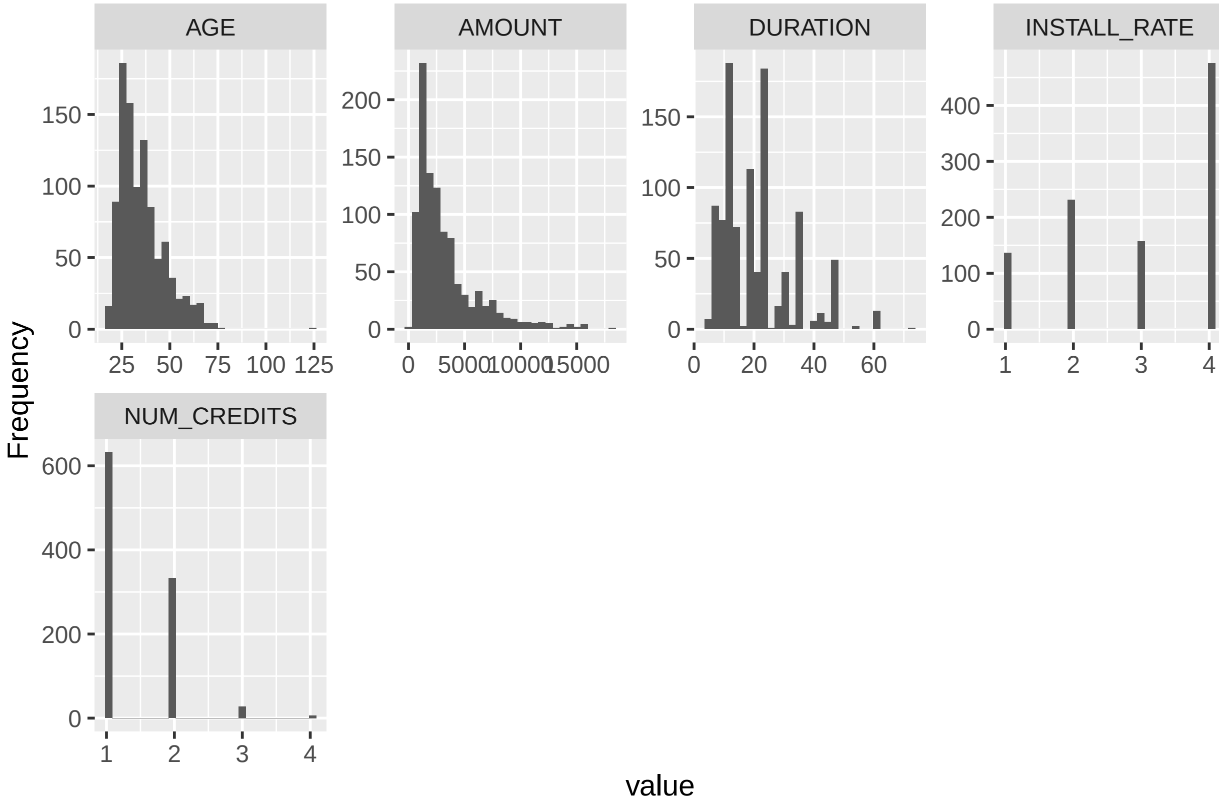
```
plot_intro(German_credit)
```



Plot the data

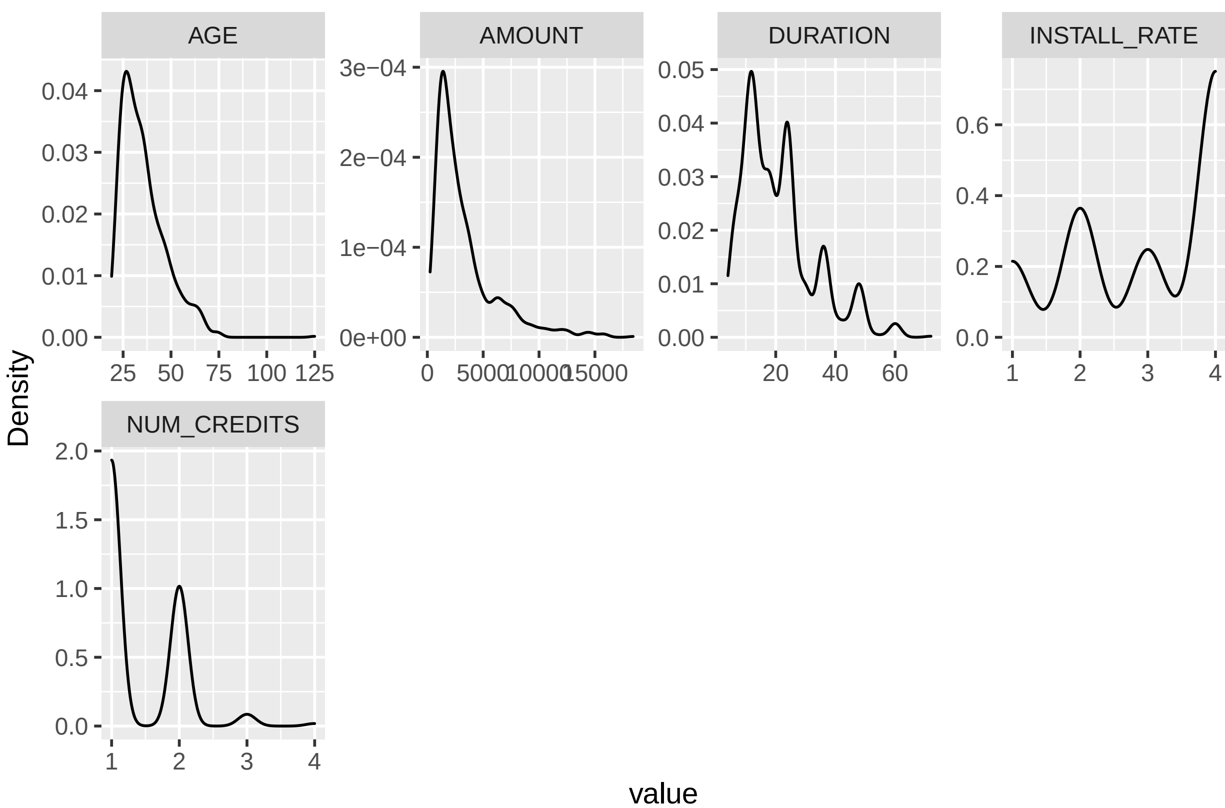
Make a plot of the continuous variables.

```
plot_histogram(German_credit)
```

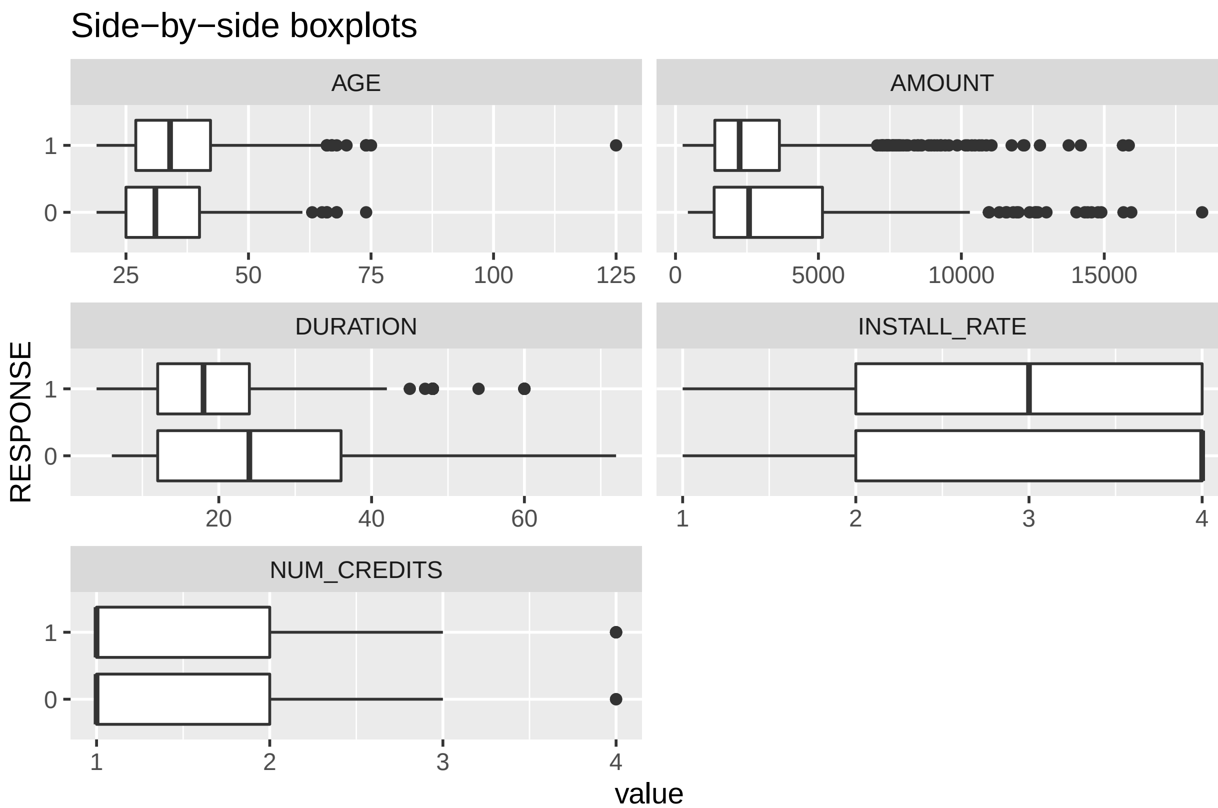


The data are not really normally distributed. They are right-tailed.

```
plot_density(German_credit)
```

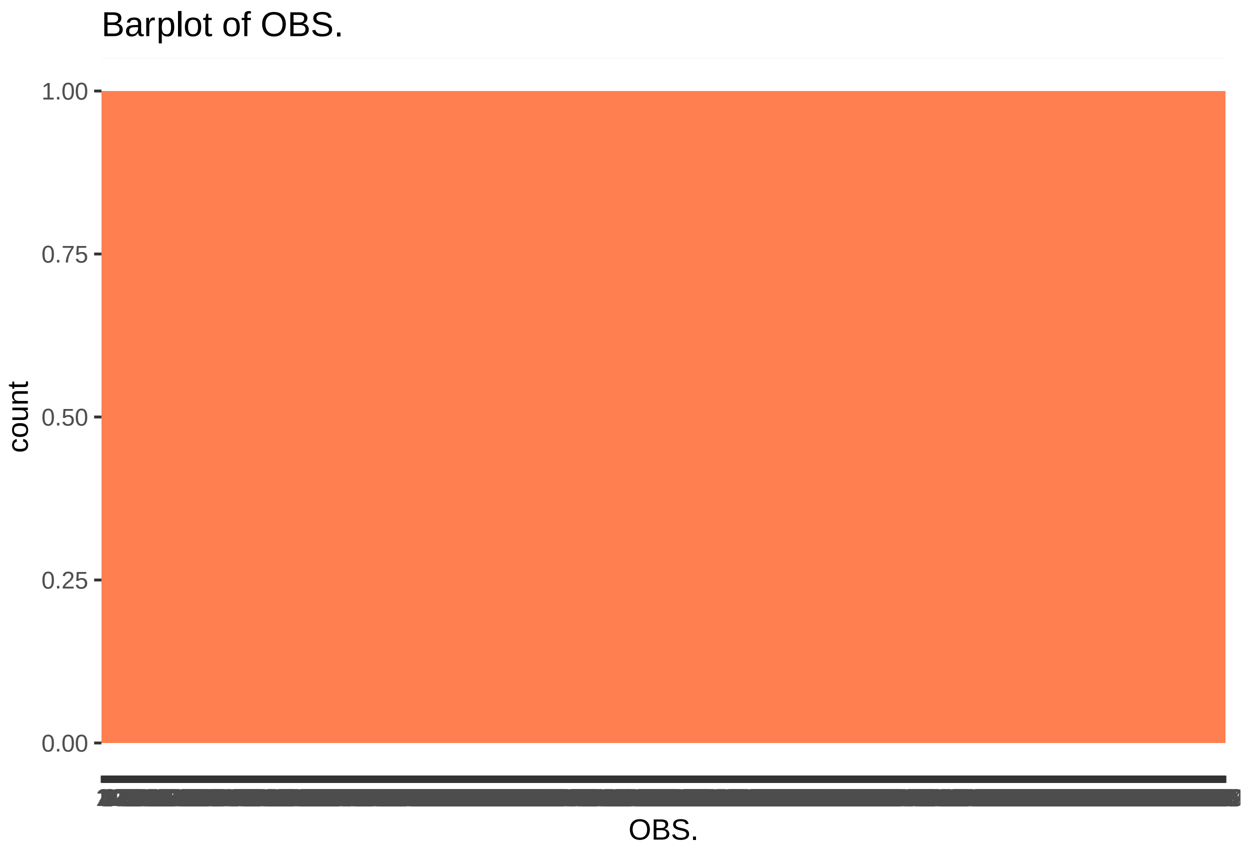


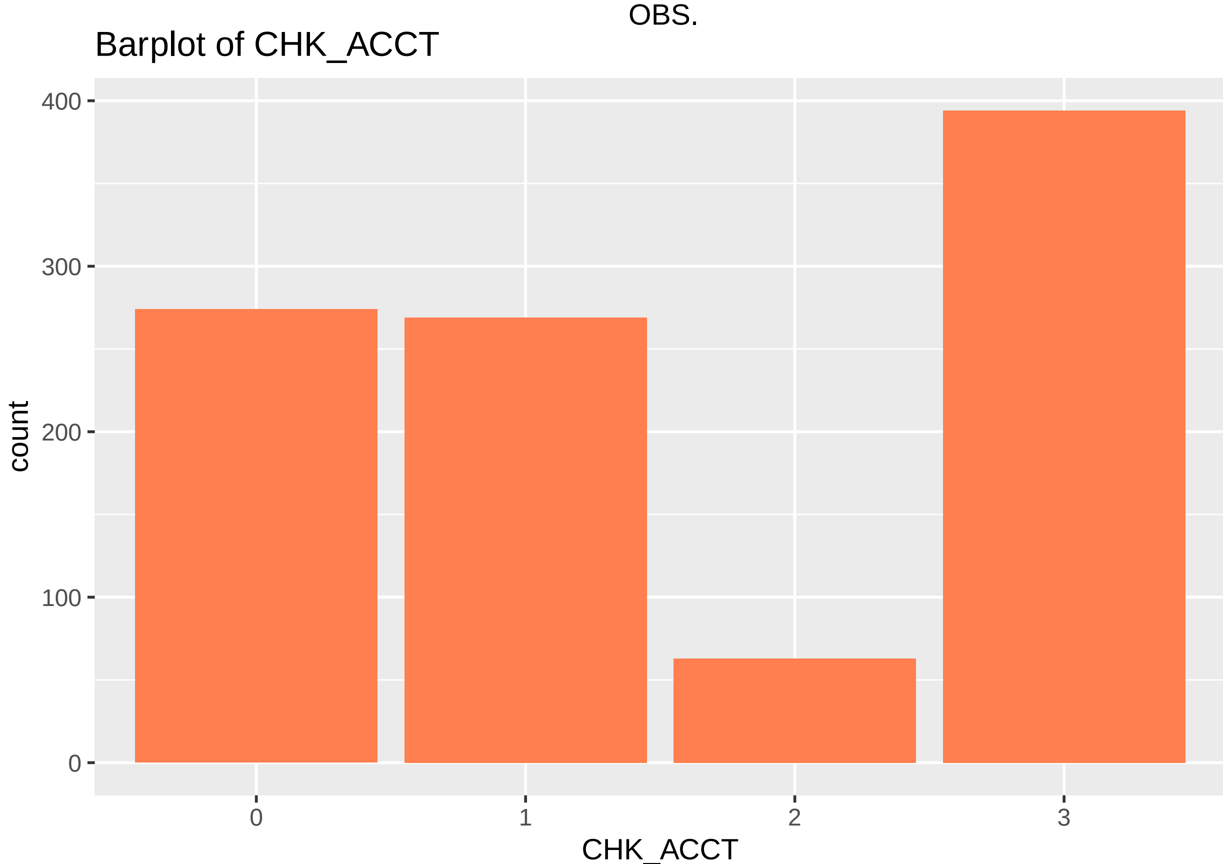
```r
plot_boxplot(German_credit, by = 'RESPONSE',  ncol = 2,
             title = "Side-by-side boxplots")
```

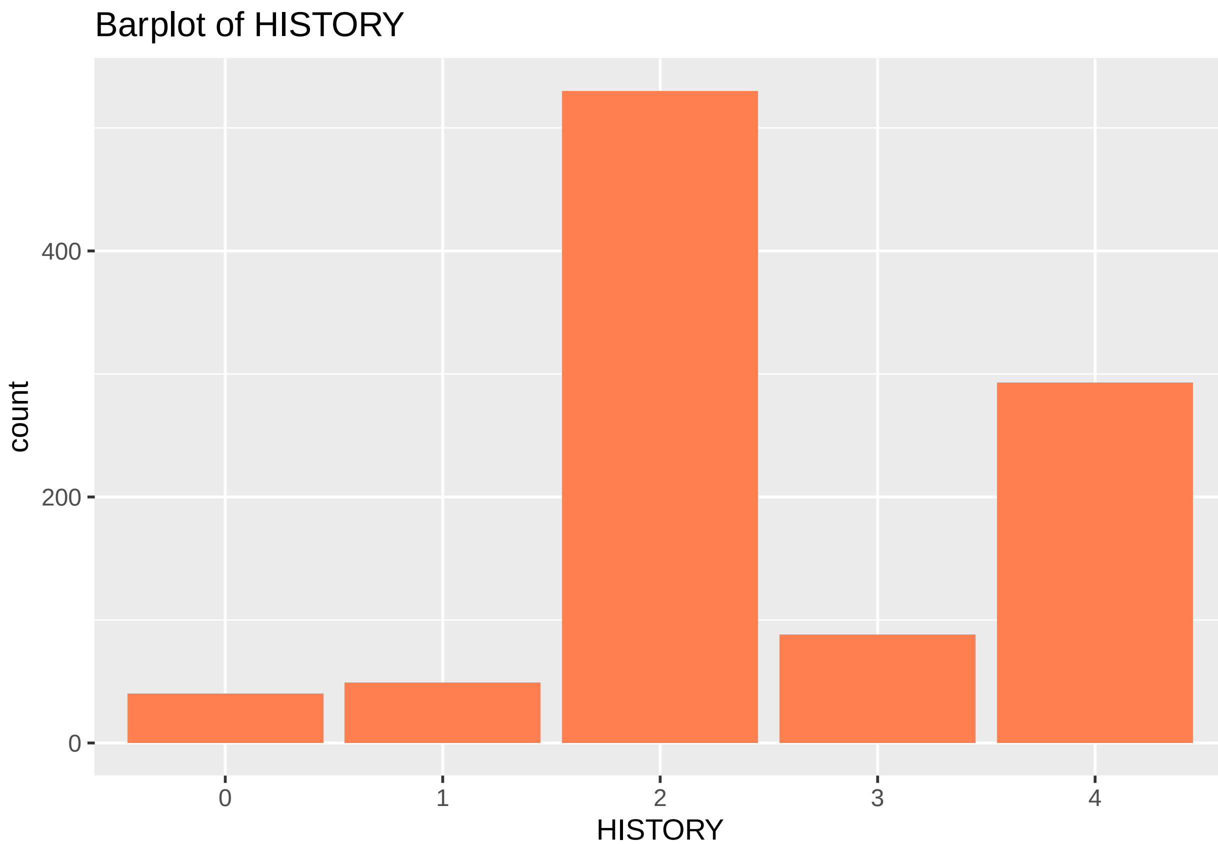


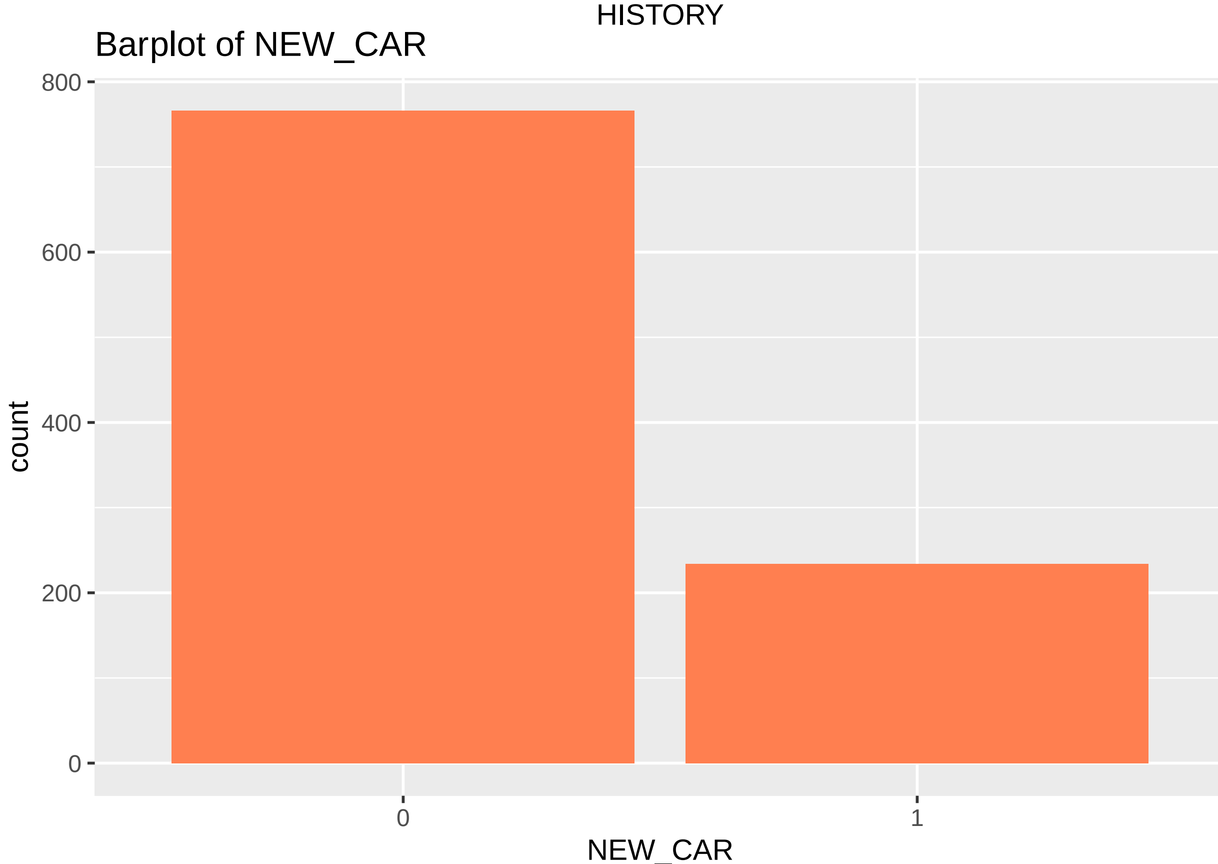
Make a barplot of the categorical variables :

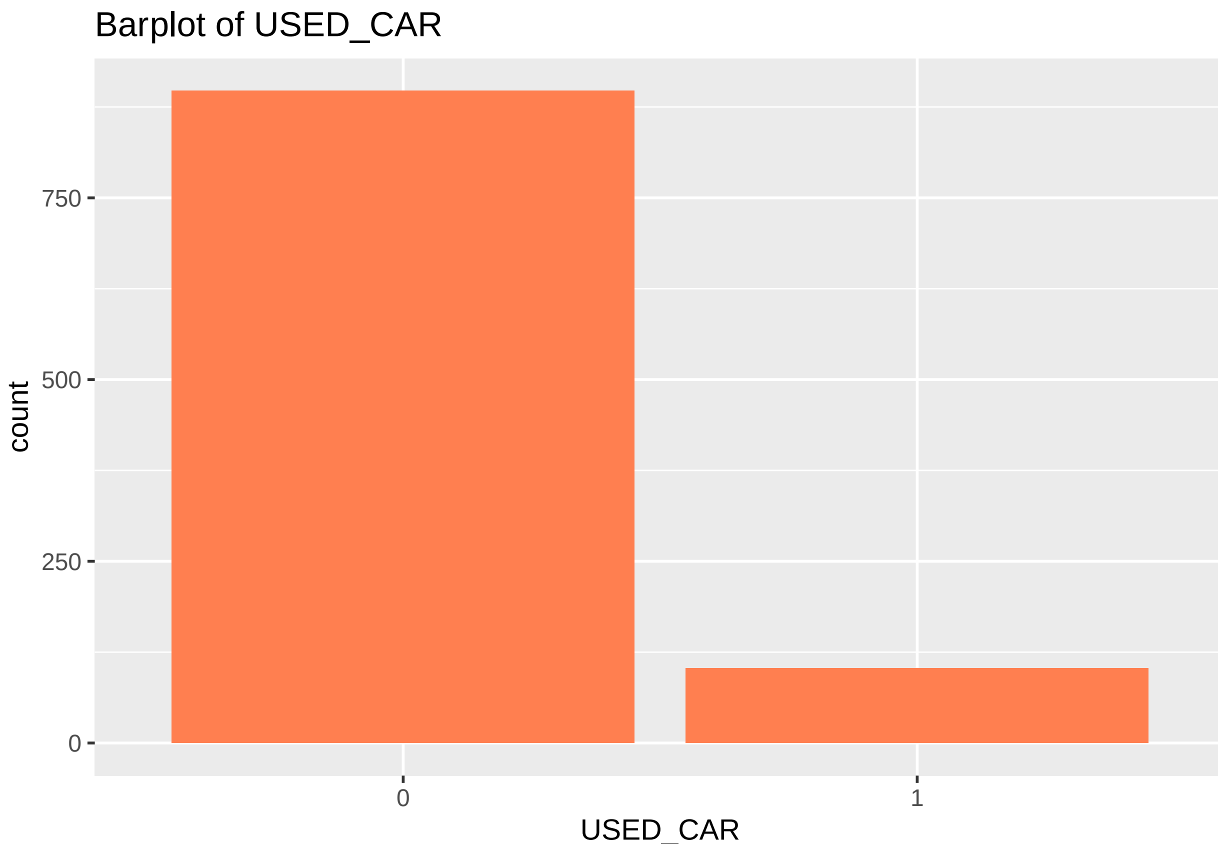
```r
# categorical_variables <- German_credit%>%
#   select(-c("DURATION", "AMOUNT", "INSTALL_RATE", "AGE", "NUM_CREDITS", "NUM_DEPENDENTS"))
# names_categorical_variables <- colnames(categorical_variables)


for (i in 1:ncol(German_credit)){
  if (class(German_credit[,i])=="factor"){
    print(ggplot(German_credit) +
            geom_bar(aes(x=German_credit[,i]), fill = "coral") +
            ggtitle(paste("Barplot of", colnames(German_credit)[i])) +
            labs(x = paste(colnames(German_credit)[i]))
            )
    }
}
```

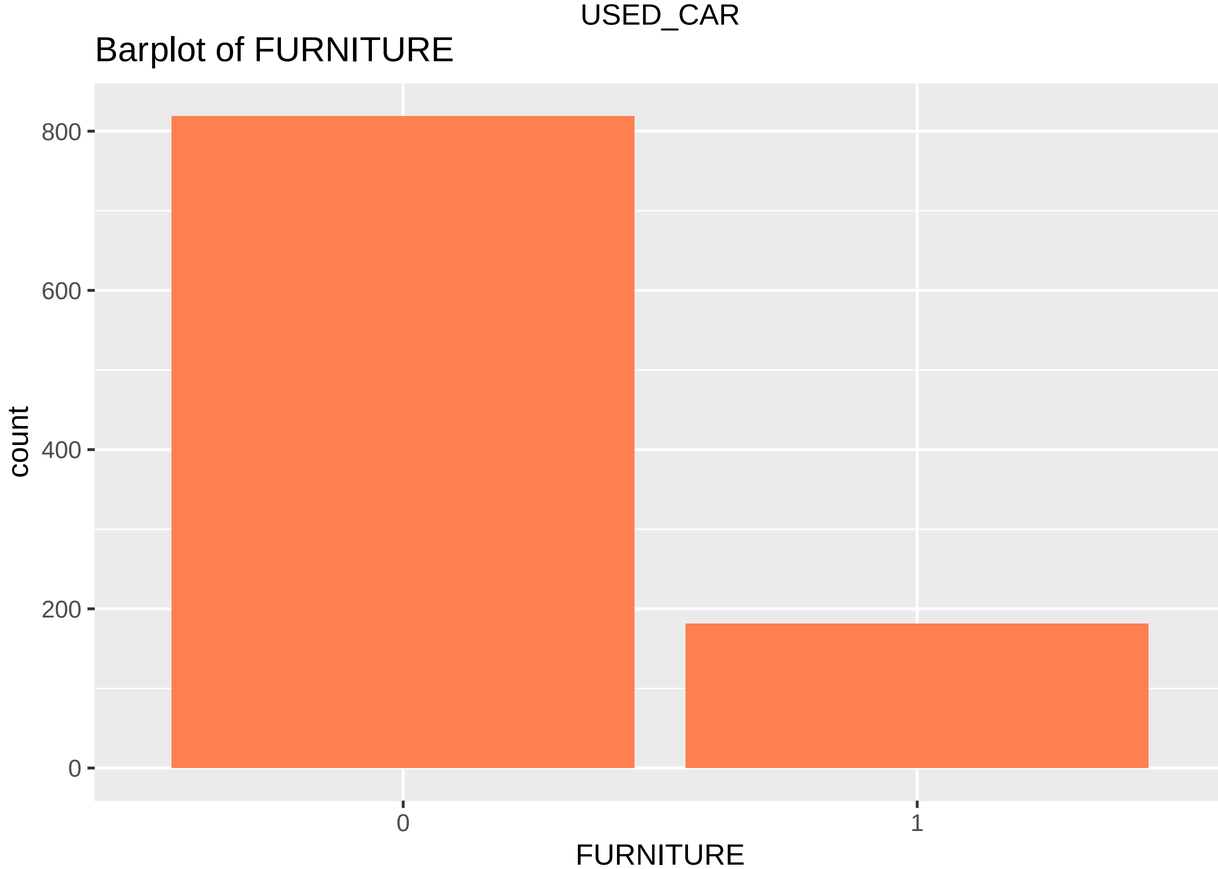

Barplot of OBS.
1.00
0.75
0.50
0.25
0.00
count
OBS.

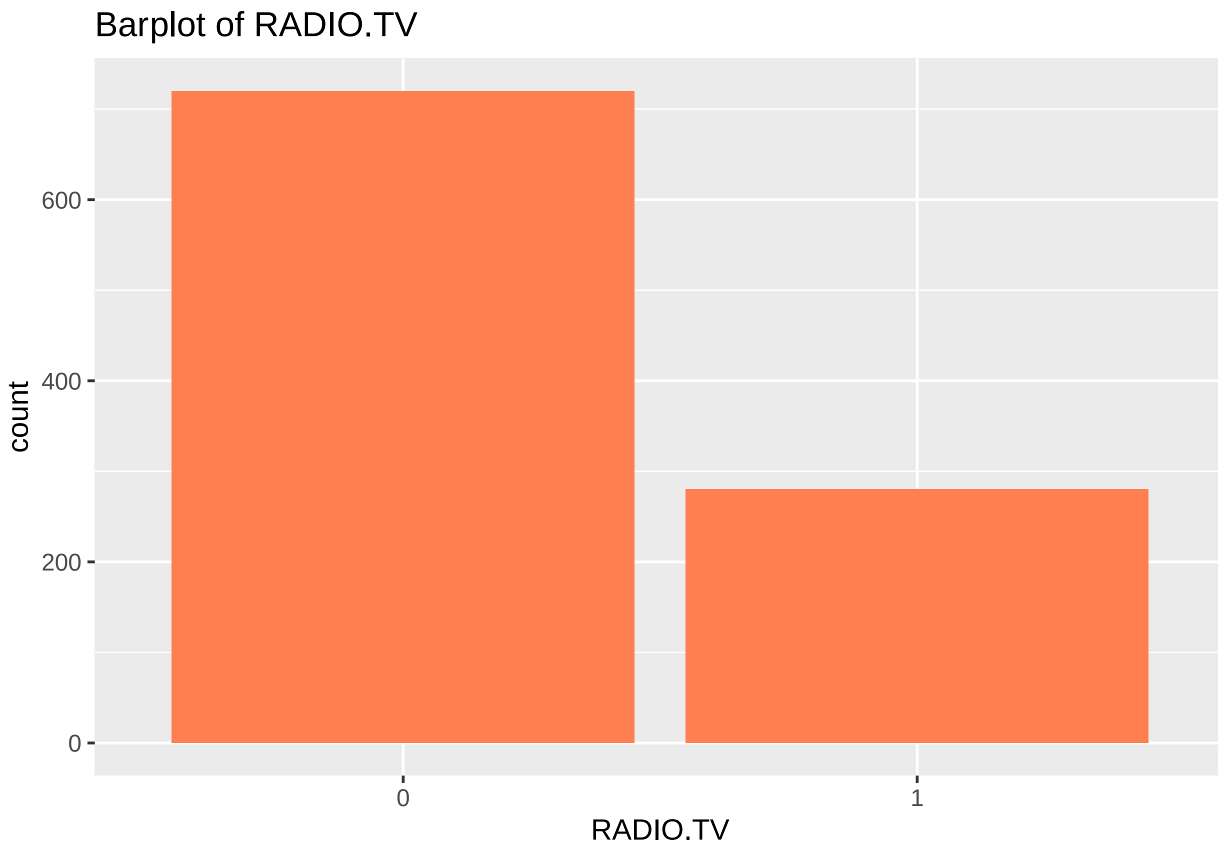

Barplot of CHK_ACCT
400
300
200
100
0
count
0
1
2
3
CHK_ACCT

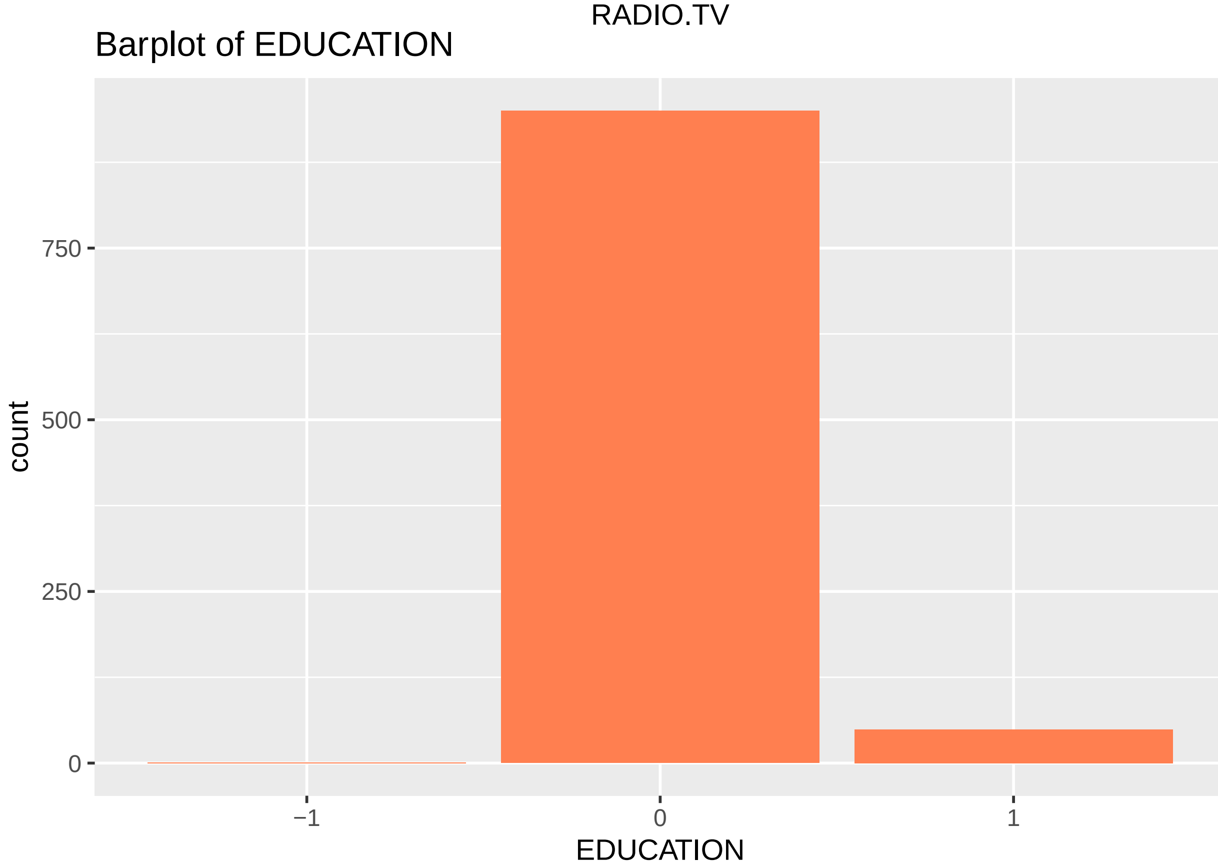

Barplot of HISTORY
count
400
200
0
0
1
2
3
4
HISTORY

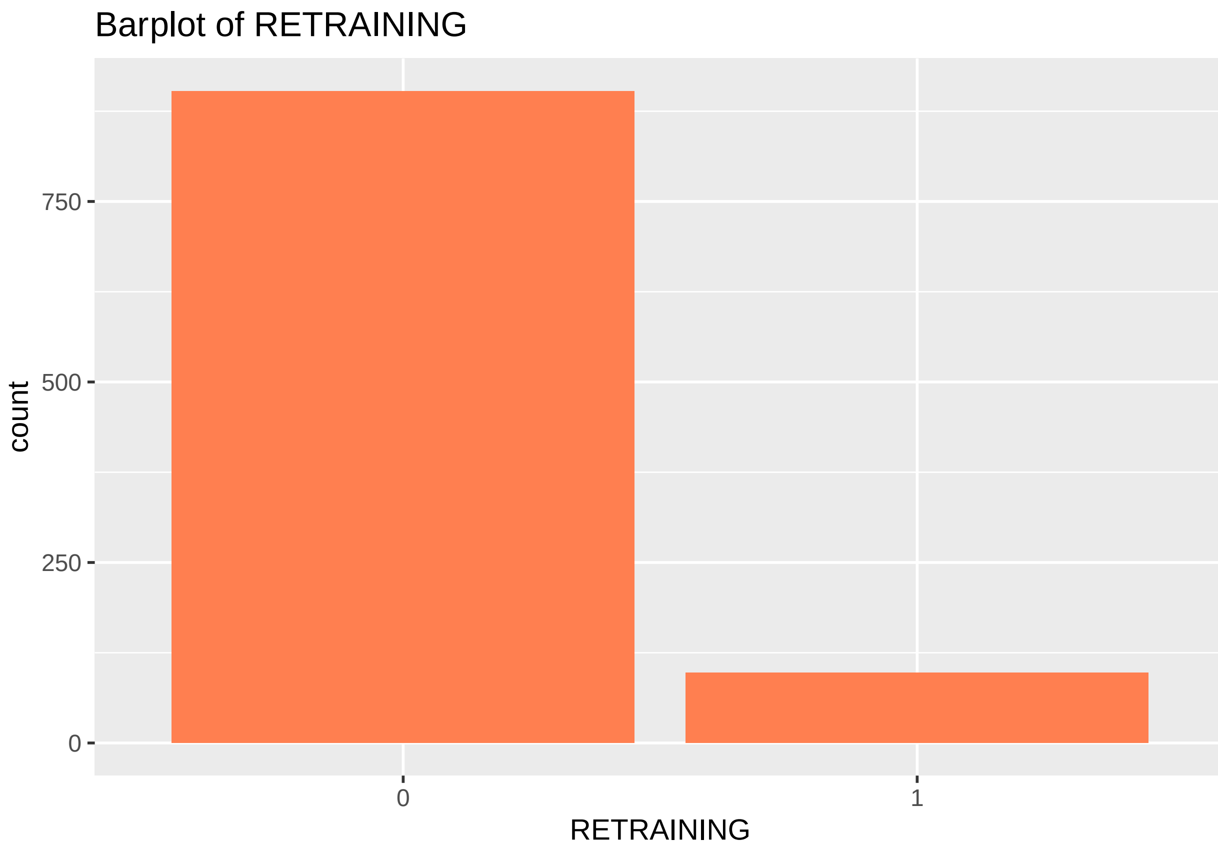

Barplot of NEW_CAR
count
800
600
400
200
0
0
1
NEW_CAR

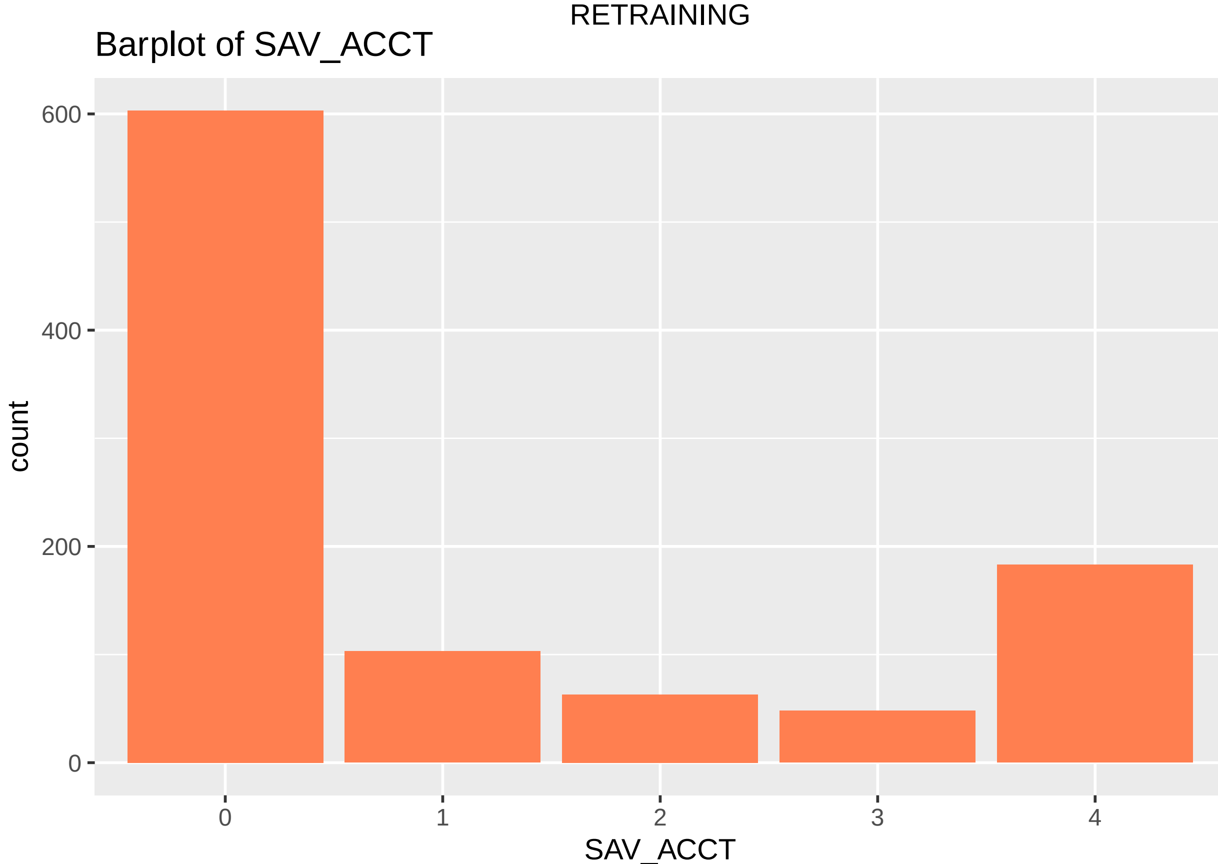

Barplot of USED_CAR
count
750
500
250
0
0
1
USED_CAR

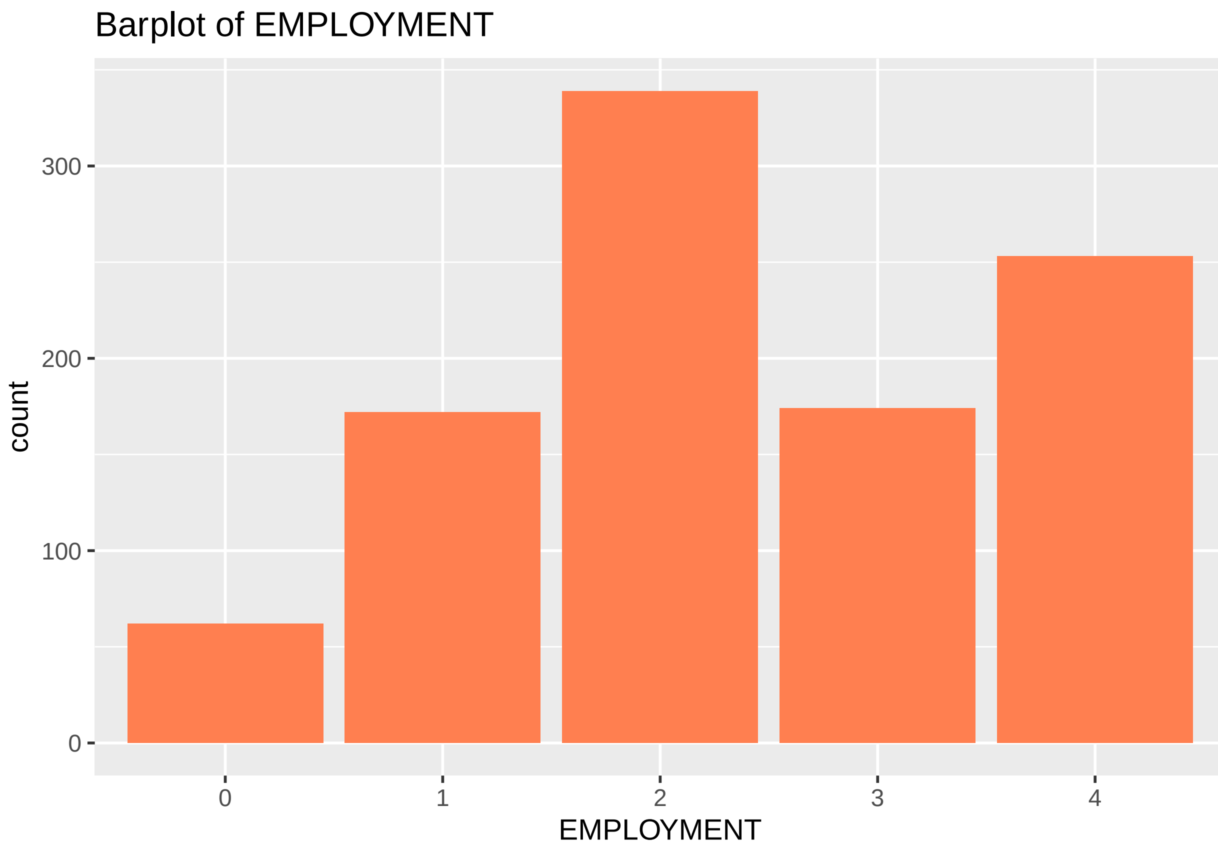

Barplot of FURNITURE
count
800
600
400
200
0
0
1
FURNITURE


Barplot of RADIO.TV
count
600
400
200
0
0
1
RADIO.TV

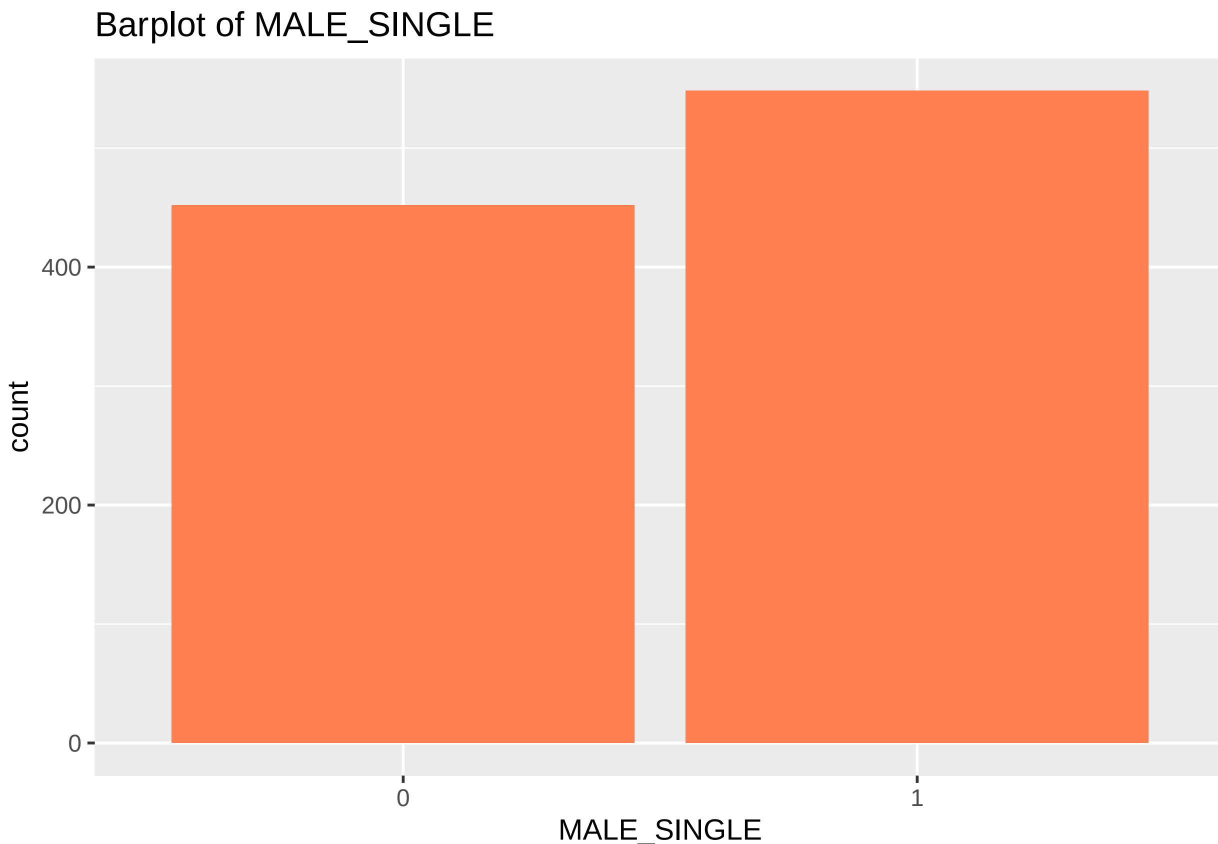

Barplot of EDUCATION
count
750
500
250
0
−1
0
1
EDUCATION

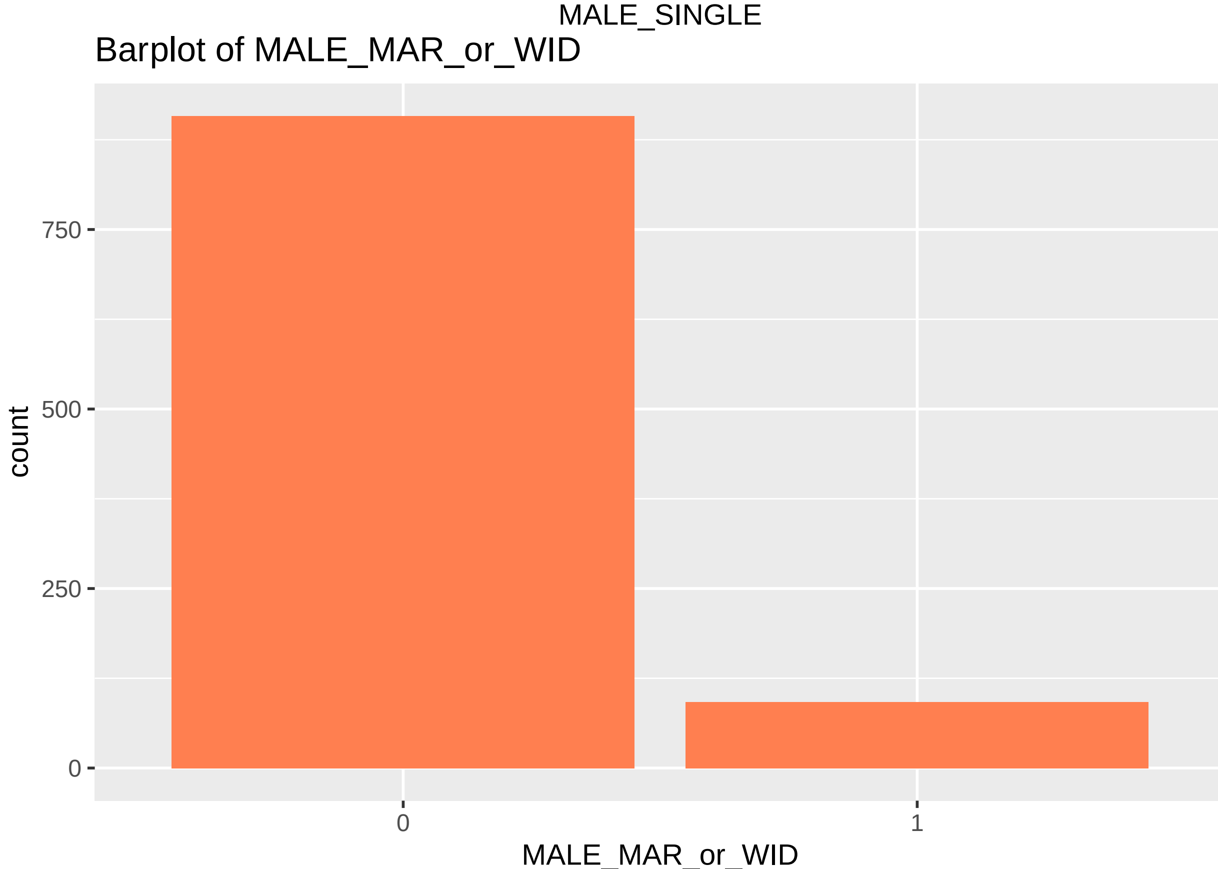

Barplot of RETRAINING
count
750
500
250
0
0
1
RETRAINING

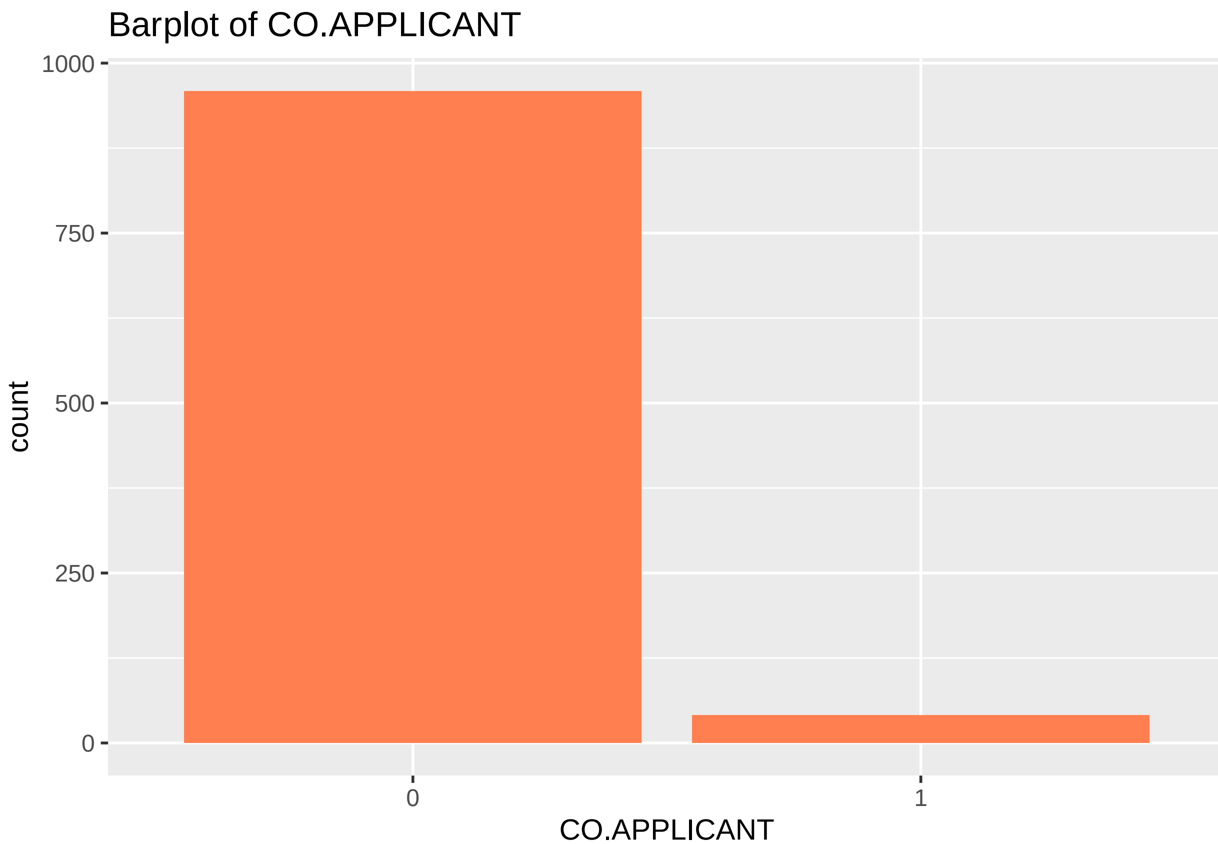

Barplot of SAV_ACCT
count
600
400
200
0
0
1
2
3
4
SAV_ACCT

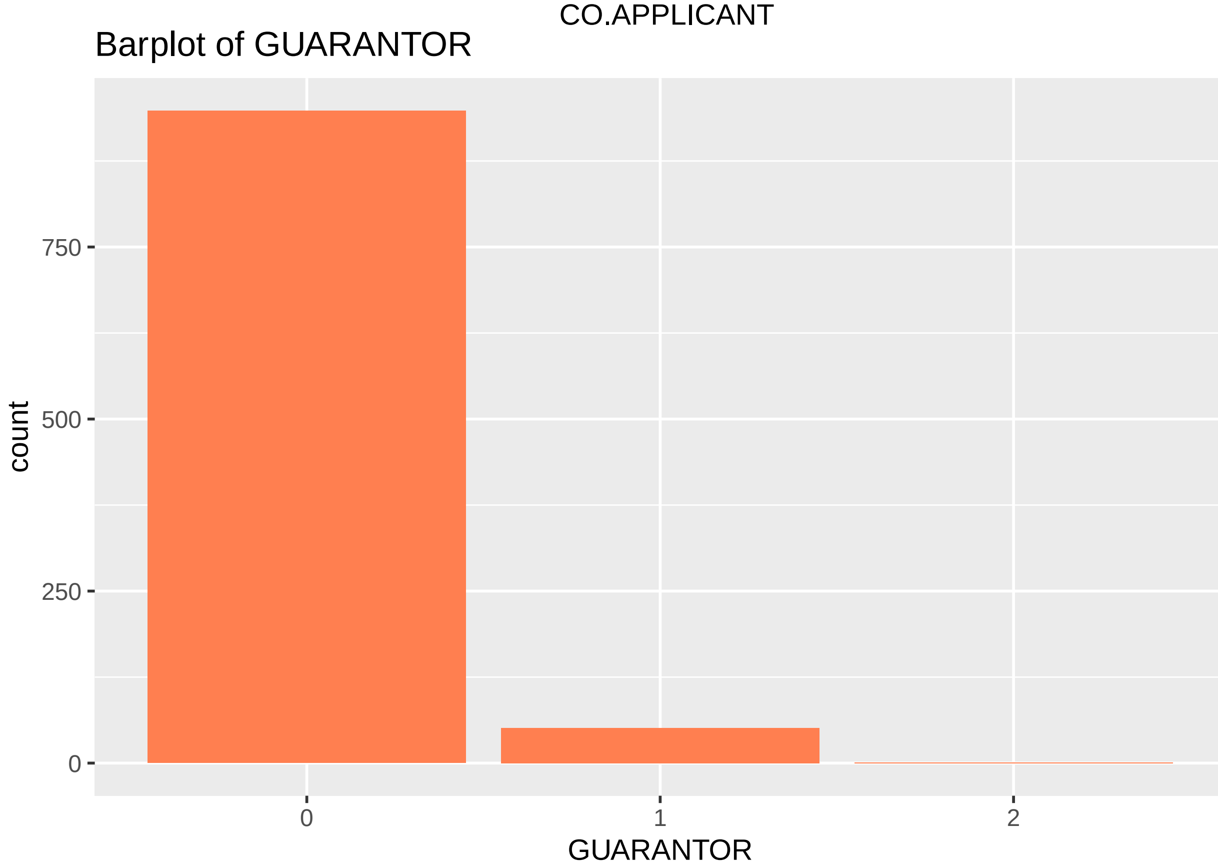

Barplot of EMPLOYMENT
count
300
200
100
0
0
1
2
3
4
EMPLOYMENT


Barplot of MALE_DIV
count
750
500
250
0
0
1
MALE_DIV

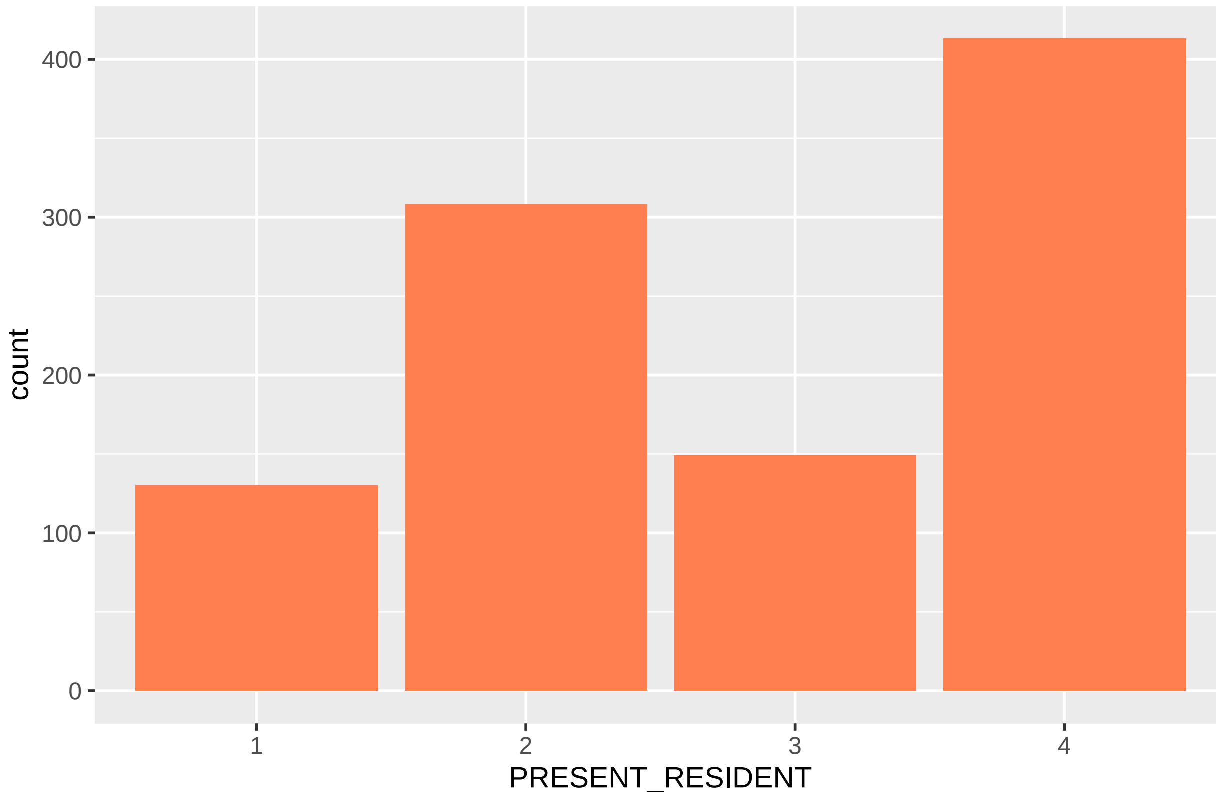

Barplot of MALE_SINGLE
count
400
200
0
0
1
MALE_SINGLE


Barplot of MALE_MAR_or_WID
count
750
500
250
0
0
1
MALE_MAR_or_WID

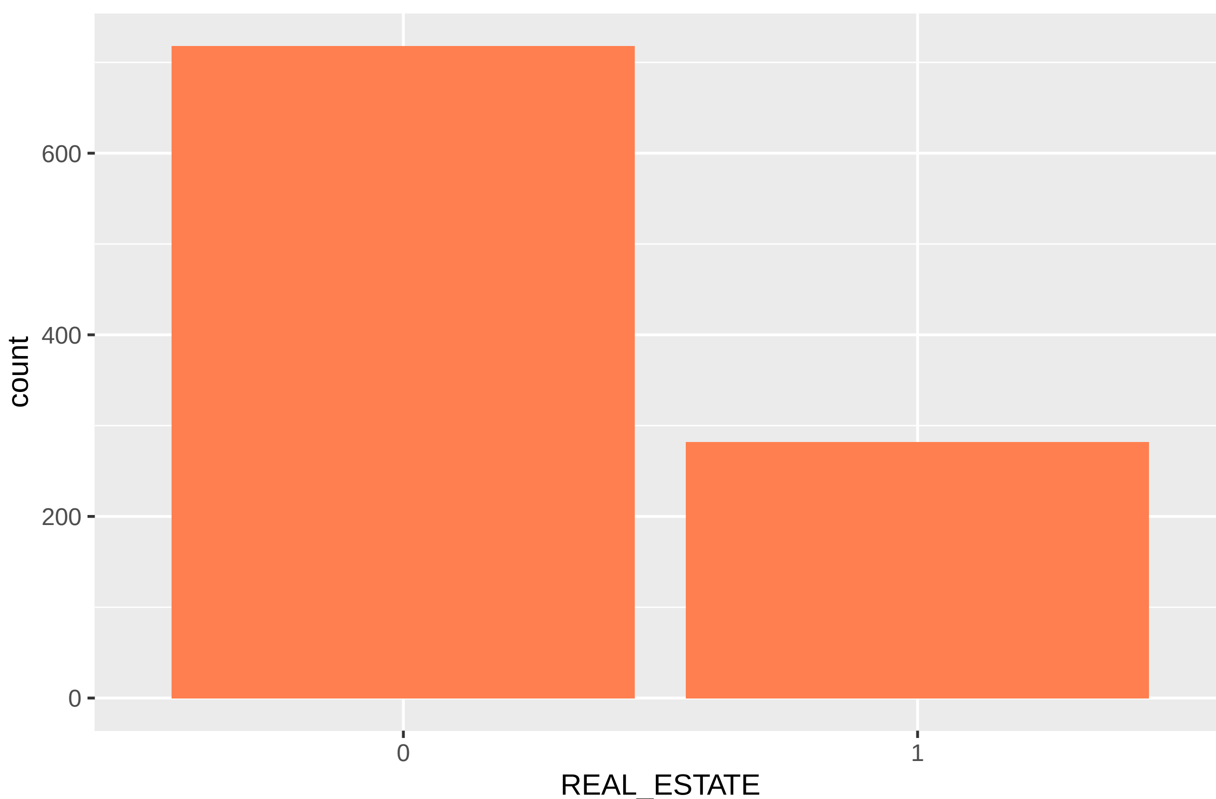

Barplot of CO.APPLICANT
1000
750
500
250
0
count
0
1
CO.APPLICANT

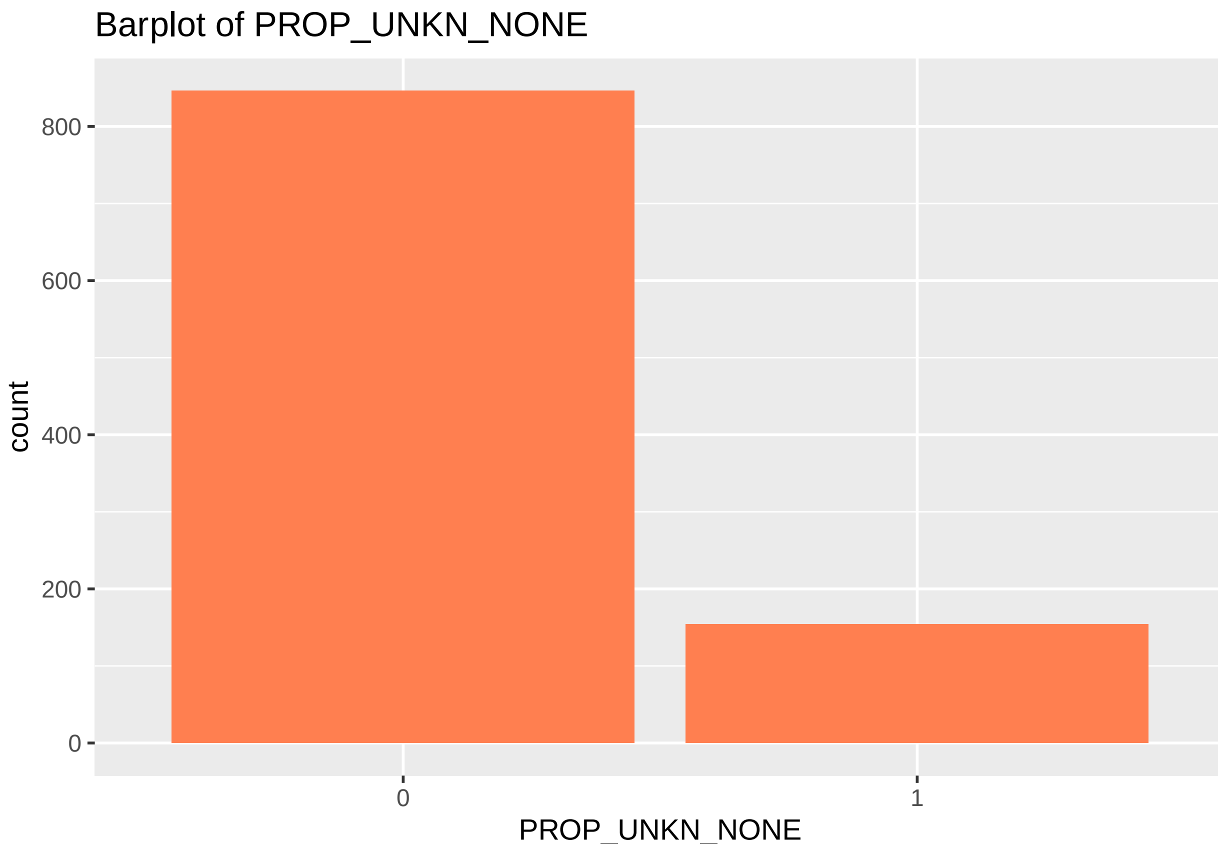

Barplot of GUARANTOR
750
500
250
0
count
0
1
2
GUARANTOR

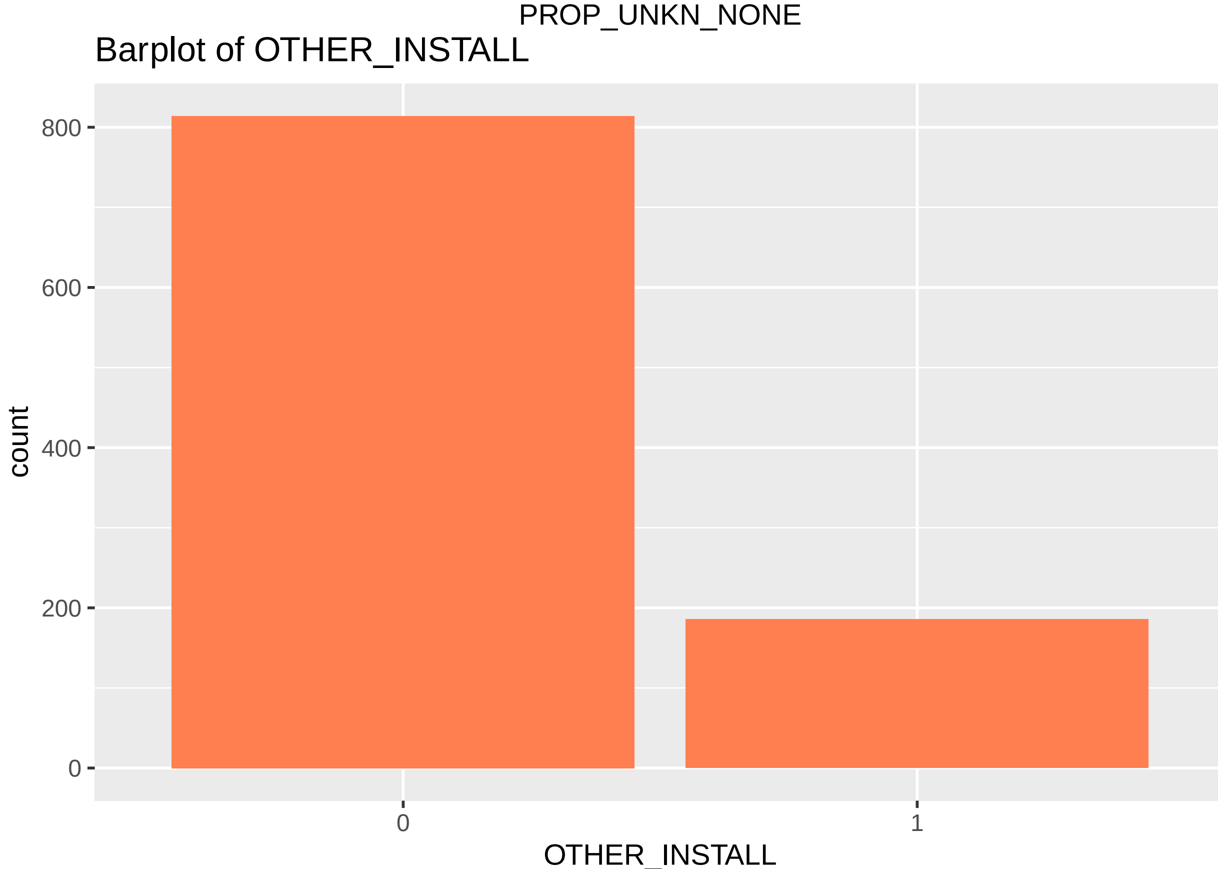
Barplot of PRESENT_RESIDENT

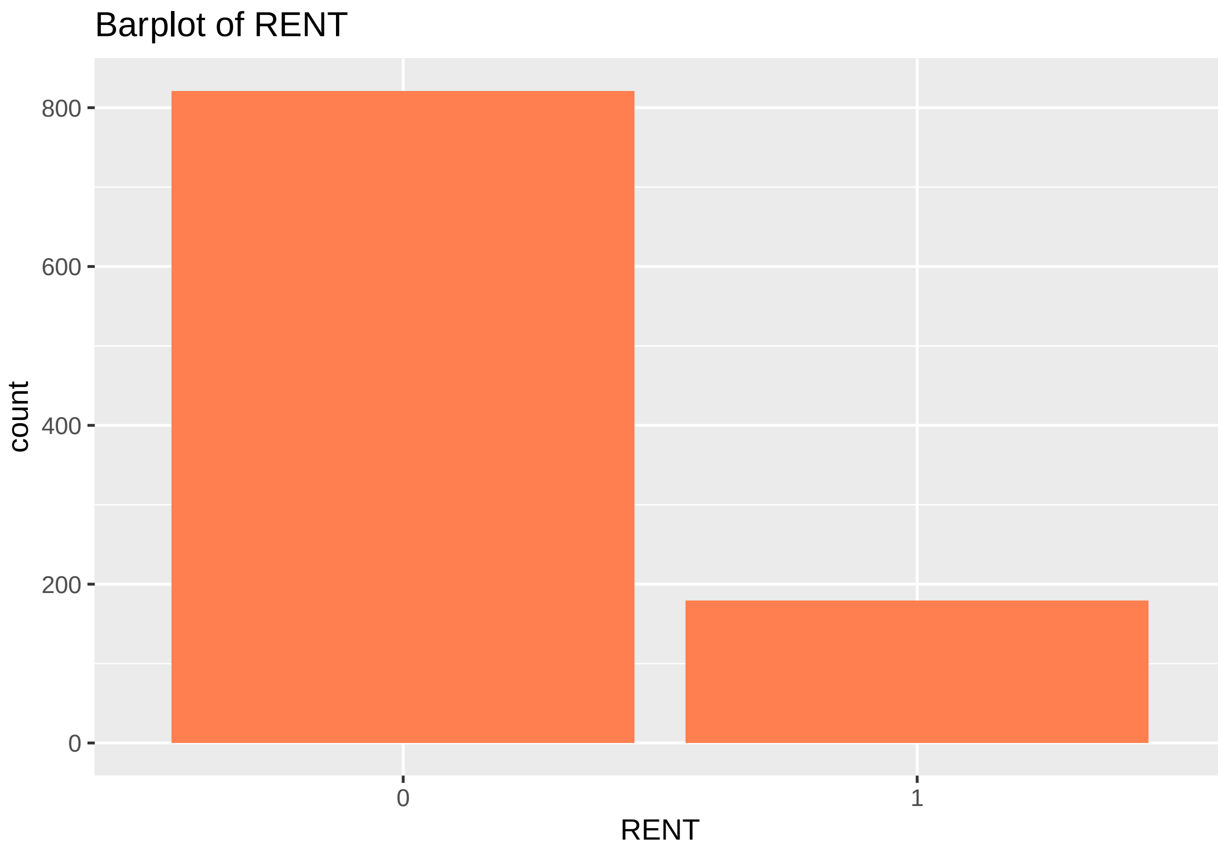


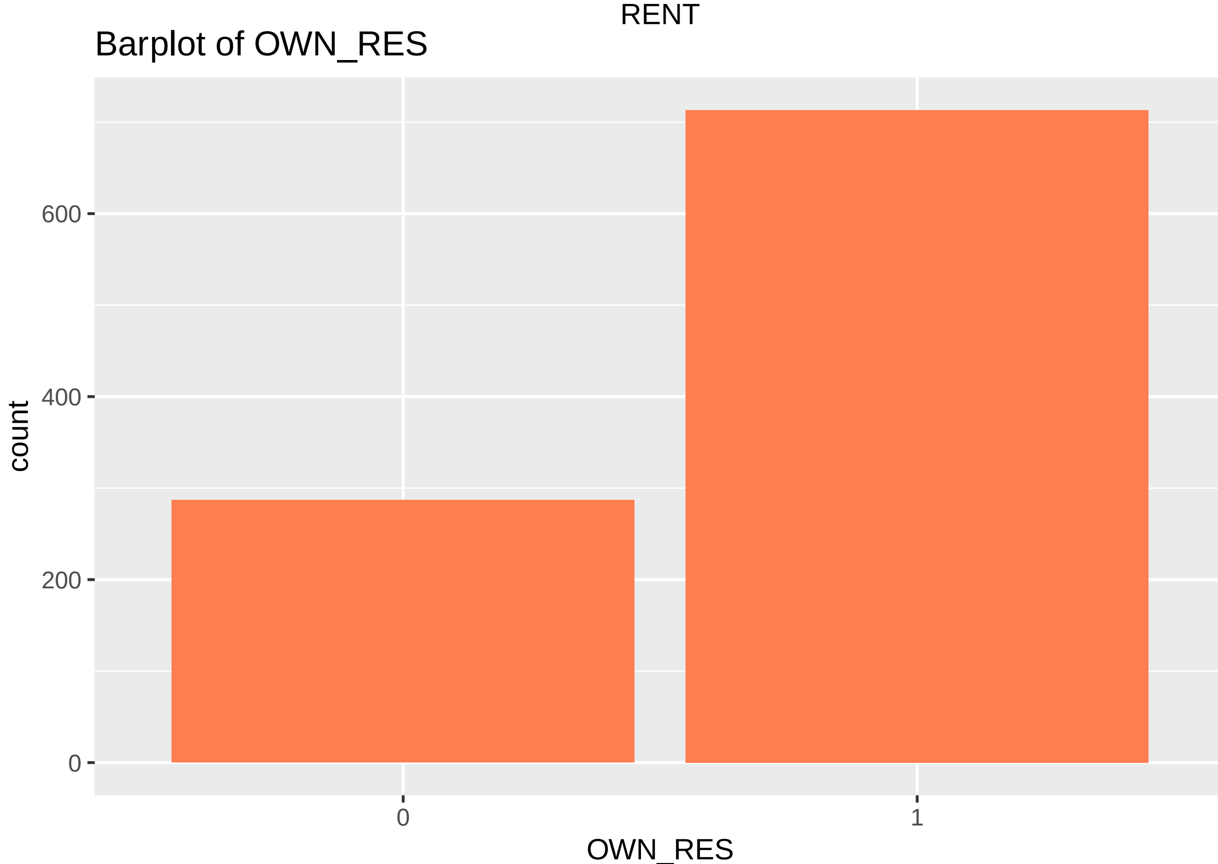
Barplot of REAL_ESTATE

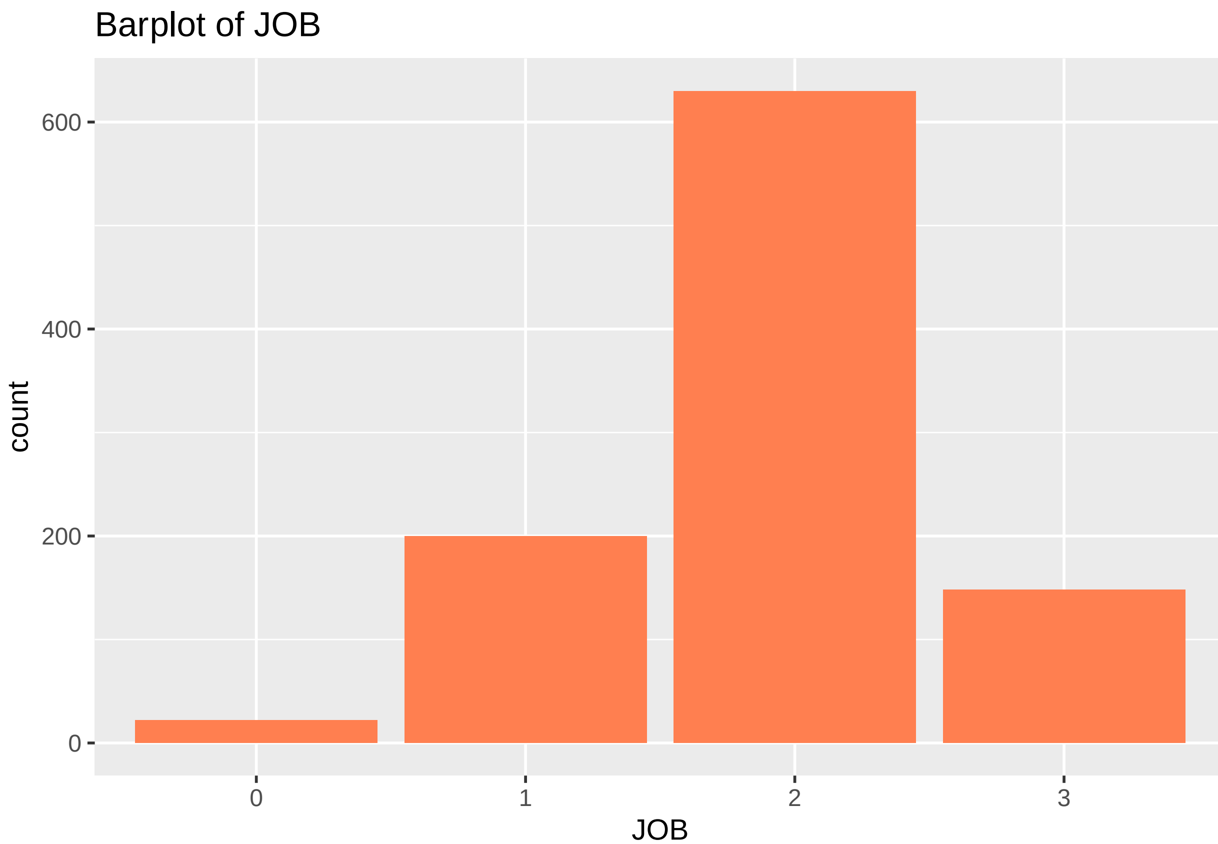


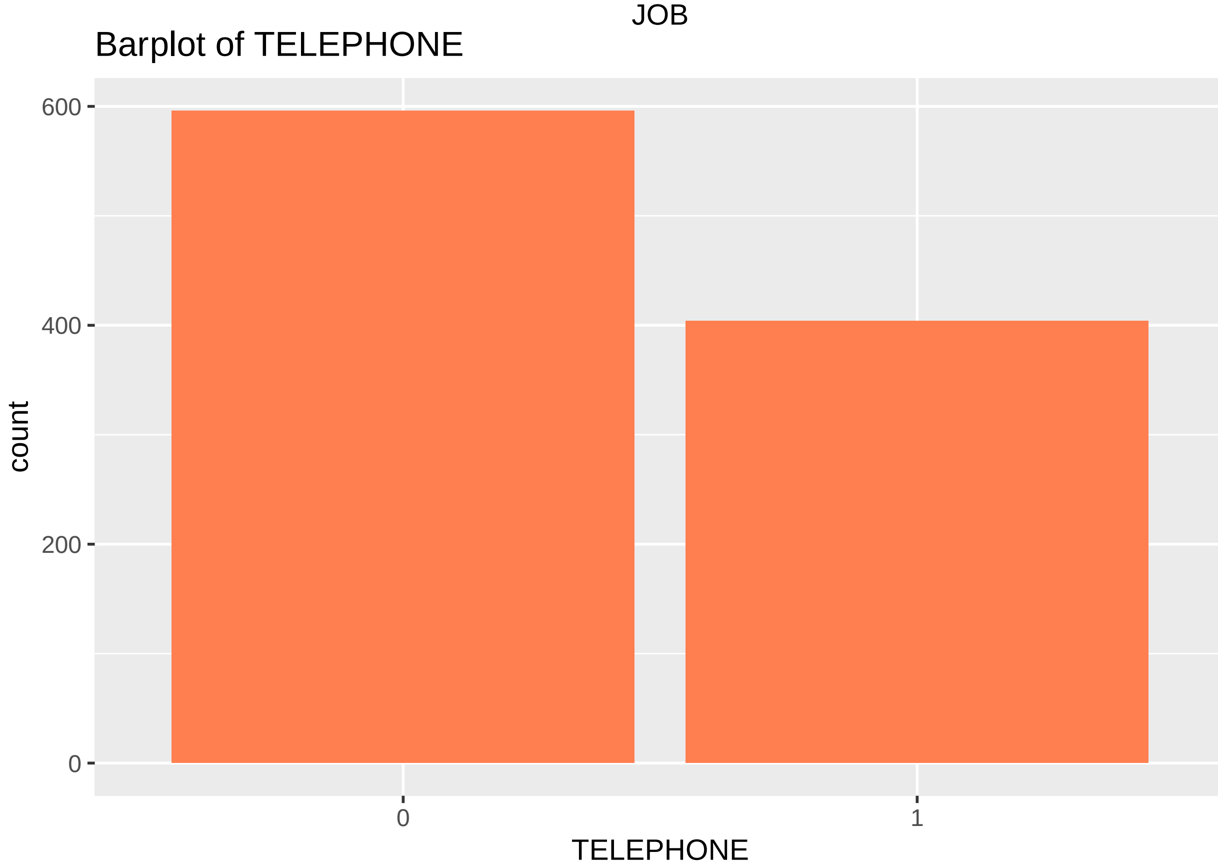

Barplot of PROP_UNKN_NONE
count
800
600
400
200
0
0
1
PROP_UNKN_NONE

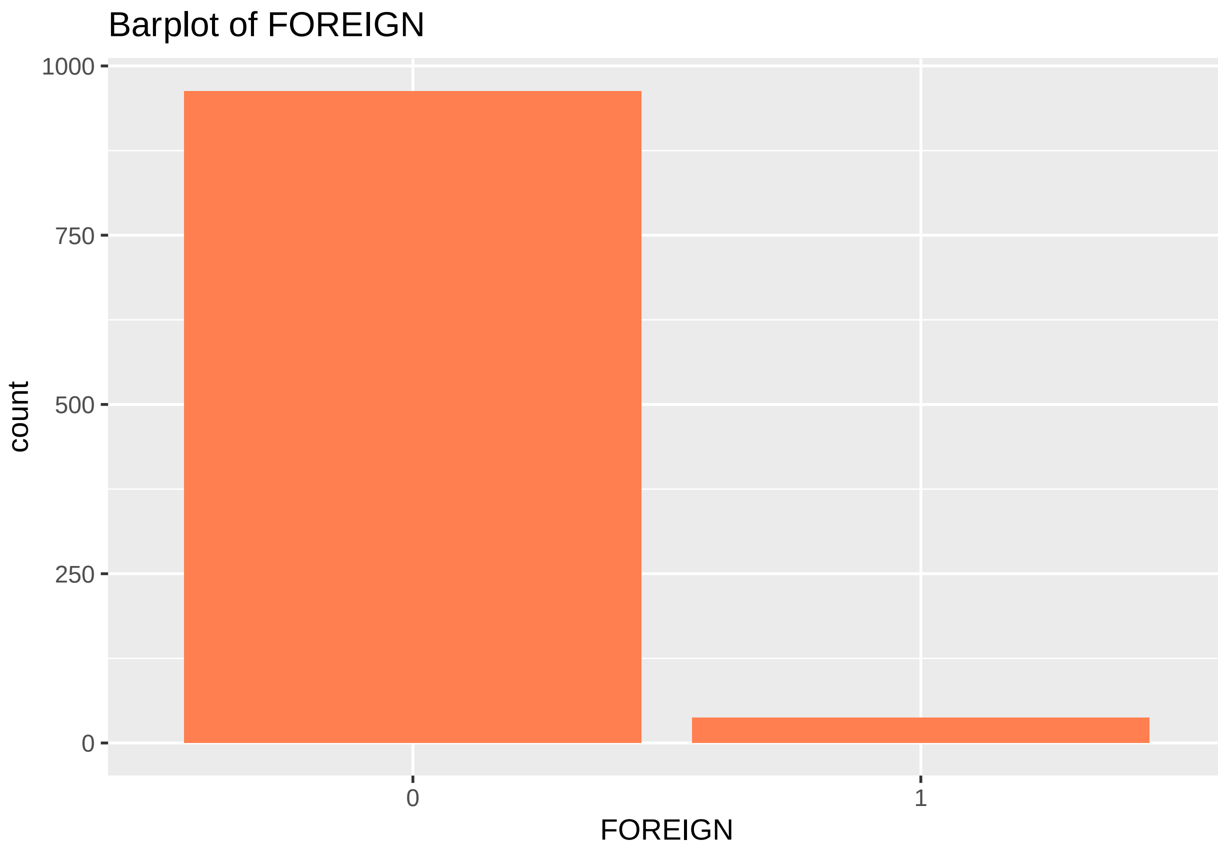

Barplot of OTHER_INSTALL
count
800
600
400
200
0
0
1
OTHER_INSTALL

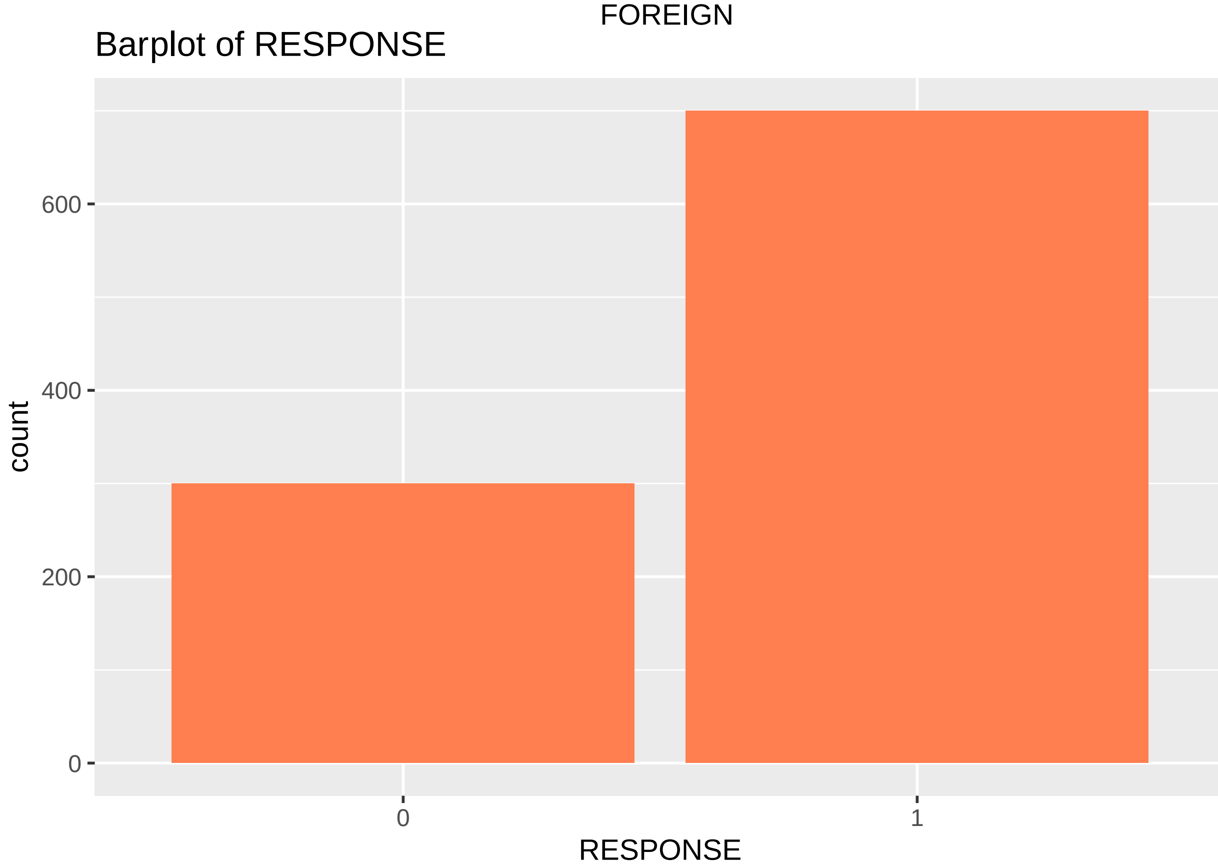

Barplot of RENT
count
800
600
400
200
0
0
1
RENT


Barplot of OWN_RES
count
600
400
200
0
0
1
OWN_RES


Barplot of JOB
count
600
400
200
0
0
1
2
3
JOB


Barplot of TELEPHONE
count
600
400
200
0
0
1
TELEPHONE





From those barplots we can see:

- The majority of people do not check their account status. (CHK_ACCT)
- Most people have an average balance of less than < 100 DM in their saving account (SAV_ACCT)
- Most of the applicants has its own residence (OWN_RES)
- Almost none of the applicants is a foreign worker (FOREIGN)

Correlation plot :

Correlation plot between continuous variables :

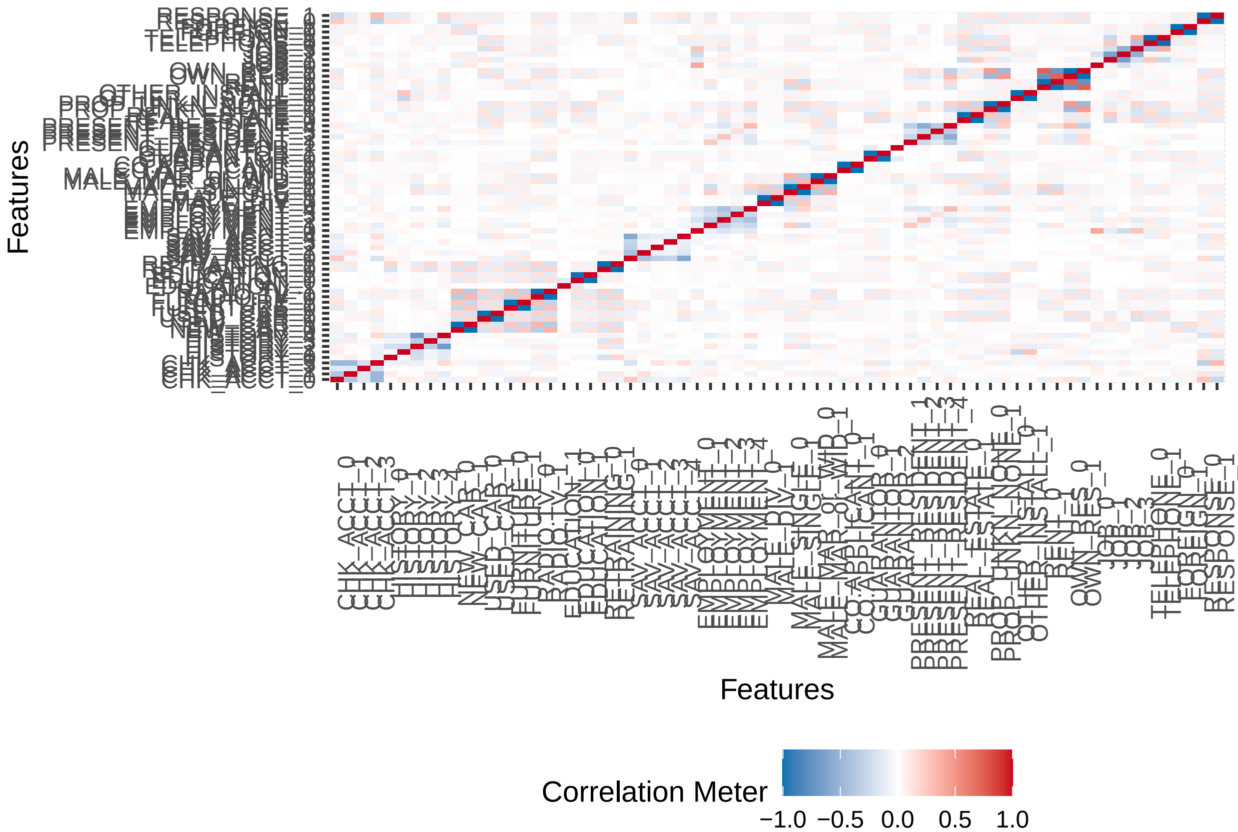
```
plot_correlation(German_credit, type= 'c', cor_args = list( 'use' = 'complete.obs'))
```

There are little correlation between the continuous variables. We can notice that there is a correlation of 62% between the variable **DURATION** and **AMOOUNT**.

Correlation plot between categorical variables :

```
plot_correlation(German_credit, type= 'd')
```

```
## 1 features with more than 20 categories ignored!
## OBS.: 1000 categories
```



It is difficult to look at the correlation since there are a lot of variables on the graph. We can still see higher correlation between **RESPONSE 1**:

- and people that do not check their account (CHK_ACCT_3)
- and people that have a critical historical account (HISTORY_4)
- and the variable *REAL_ESTATE* (REAL_ESTATE)
- and applicant that does not have their own property (PROP_UNKN_NONE_0)
- and applicant that have their own residence (OWN_RES_1)

We can also see some correlation between **RESPONSE 0**:

- and people that have a checking account status < 0 DM (CHK_ACCT_0)
- and people that have an average balance in savings account < 100 DM (SAV_ACCT_0)
- and the variable *REAL_ESTATE* (REAL_ESTATE)

Fitting a model :

Let's try a lassification tree

```
german.ct <- rpart(RESPONSE ~ ., method = "class", data = German_credit)
summary(german.ct)
```

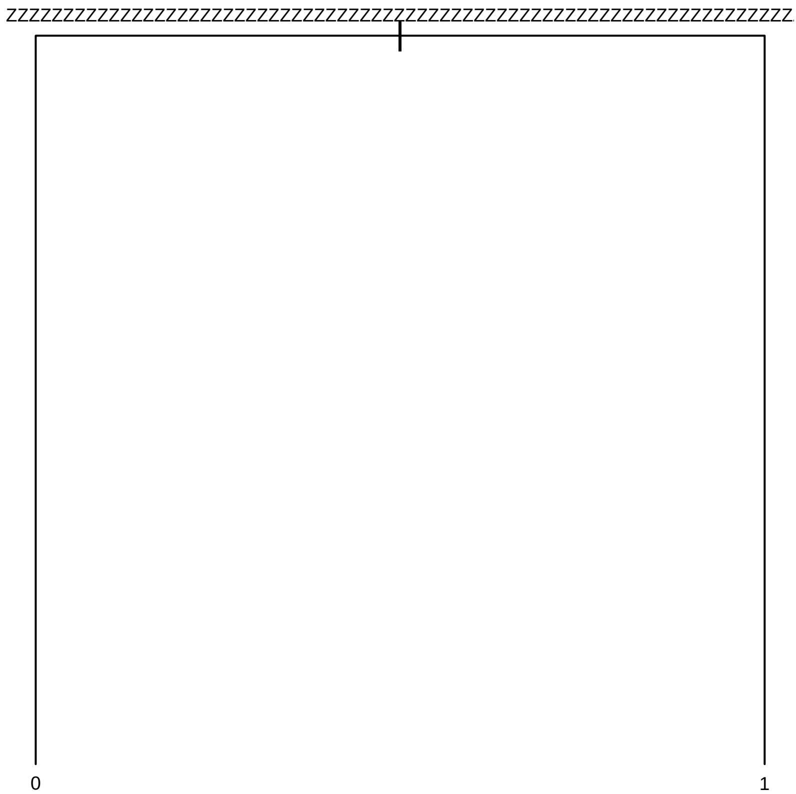
```
## Call:
## rpart(formula = RESPONSE ~ ., data = German_credit, method = "class")
##   n= 1000
##
##     CP nsplit rel error    xerror       xstd
## 1 1.00      0         1 1.0000000 0.04830459
## 2 0.01      1         0 0.9733333 0.04792772
##
```

```
## Variable importance
##     OBS.  HISTORY   AMOUNT DURATION
##       88        5        4        3
##
## Node number 1: 1000 observations,    complexity param=1
##   predicted class=1  expected loss=0.3  P(node) =1
##     class counts:   300   700
##    probabilities: 0.300 0.700
##   left son=2 (300 obs) right son=3 (700 obs)
##   Primary splits:
##       OBS.     splits as  RLRRLRRRRLLLRLRLRRLRRRRRRRRRRLRRRRRLRLRRRRRRLRRRRRRRRRLRLRRLRRLLRRRRLRRRRRR
##       CHK_ACCT splits as  LLRR, improve= 47.90962, (0 missing)
##       HISTORY  splits as  LLRRR, improve= 17.06212, (0 missing)
##       SAV_ACCT splits as  LLRRR, improve= 14.80642, (0 missing)
##       DURATION < 34.5  to the right, improve= 13.62155, (0 missing)
##   Surrogate splits:
##       HISTORY  splits as  LLRRR,     agree=0.717, adj=0.057, (0 split)
##       AMOUNT   < 10918 to the right, agree=0.713, adj=0.043, (0 split)
##       DURATION < 43.5  to the right, agree=0.710, adj=0.033, (0 split)
##
## Node number 2: 300 observations
##   predicted class=0  expected loss=0  P(node) =0.3
##     class counts:   300     0
##    probabilities: 1.000 0.000
##
## Node number 3: 700 observations
##   predicted class=1  expected loss=0  P(node) =0.7
##     class counts:     0   700
##    probabilities: 0.000 1.000
```

The model is not working yet.

```r
par(pty = "s", mar = c(1, 1, 1, 1))
plot(german.ct, cex = 1)
text(german.ct, cex = 0.6)
```

```
## Warning in labels.rpart(x, minlength = minlength): more than 52 levels in a
## predicting factor, truncated for printout
```


0
1