Exploratory Data Analysis

Elodie Kwan and Katia Voltz

2022-04-26

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

## Le chargement a nécessité le package : lattice

## Le chargement a nécessité le package : survival

## Le chargement a nécessité le package : Formula

##
## 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

##
## Attachement du package : 'psych'

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

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

##
## Attachement du package : 'gridExtra'

## L'objet suivant est masqué depuis 'package:dplyr':
##
##     combine

## Warning in fun(libname, pkgname): couldn't connect to display ":0"

## system might not have X11 capabilities; in case of errors when using dfSummary(), set st_options(use
##
## Attachement du package : 'summarytools'
```

```
## Les objets suivants sont masqués depuis 'package:Hmisc':
##
##     label, label<-
```

```
## L'objet suivant est masqué depuis 'package:tibble':
##
##     view
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

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

```
##
## Attachement du package : 'caret'
```

```
## L'objet suivant est masqué depuis 'package:survival':
##
##     cluster
```

```
## L'objet suivant est masqué depuis 'package:purrr':
##
##     lift
```

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

```
##
## Attachement du package : 'foreach'
```

```
## Les objets suivants sont masqués depuis 'package:purrr':
##
##     accumulate, when
```

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

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

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

In this section, we will proceed to an exploratory data analysis of the **German Credit data**.

Let's start by importing the dataset.

```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 :

```
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 ...
```

```
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
```

We notice that the variable **EDUCATION** has a minimum value of '-1' but it should be '0' since there are only 2 levels (0 and 1). Indeed, the observation 37 indicate a value of '-1' for **EDUCATION**. We notice another strange value, in the variable **GUARANTOR**, the maximum value is of '2' while it does not mean anything in our data set.

After discussion with the Banker, he gave us the correct values to these 2 mistakes. Observation 37 of **EDUCATION** and observation 234 of **GUARANTOR** should be equal to 1.

```
German_credit$EDUCATION[37] <- 1
German_credit$EDUCATION <- as.factor(German_credit$EDUCATION)

German_credit$GUARANTOR[234] <- 1
German_credit$GUARANTOR <- as.factor(German_credit$GUARANTOR)
```

We see that the variable **AGE** has a maximum of 125. This is strange because it is very unlikely that someone lives to the age of 125. We talked to the banker and he confirmed our doubts by telling us that a mistake has been made. At the observation 537, the correct age of the client is 75 years old. He asks us to correct this value in our data set.

```
German_credit$AGE[537] <- 75
```

- There are no missing values.

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

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

- The response variable is the '**Response**' variable - last column on the data.

Response variable : credit rating is good

1. 0 : No
2. 1 : Yes

We have to make sure that the class of the variables are correct. As described above, all the variables are defined as *integer* but we know that we should have numerical and categorical variables in our dataset. Therefore, we have to transform the class of some of them.

```
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/ 2 levels "0","1": 1 1 2 1 1 2 1 1 1 1 ...
##  $ 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/ 2 levels "0","1": 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 ...
```

The binomial data are set as factors and the others as numerical.

We can now describe the variables one more time and we should get better results.

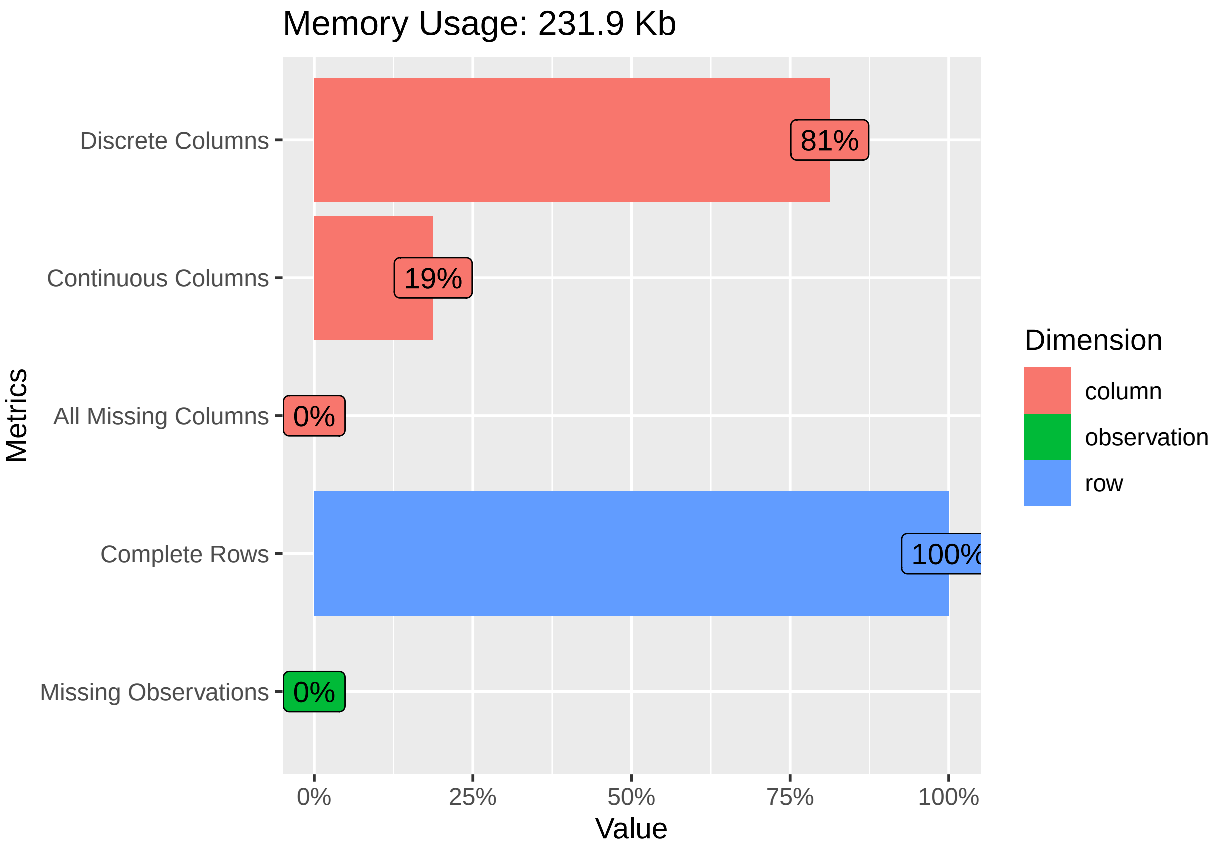
```
describe(German_credit)
```

```
##                   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    1.05    0.22    1.0    1.00    0.00   1     2
## 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    0.22    1.0    1.00    0.00   1     2
## 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   35.55   11.38   33.0   34.17   10.38  19    75
## 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*            1  4.12    15.02  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*            1  4.03    14.25  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                  56  1.02     0.58  0.36
## 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       237424
```

```
plot_intro(German_credit)
```

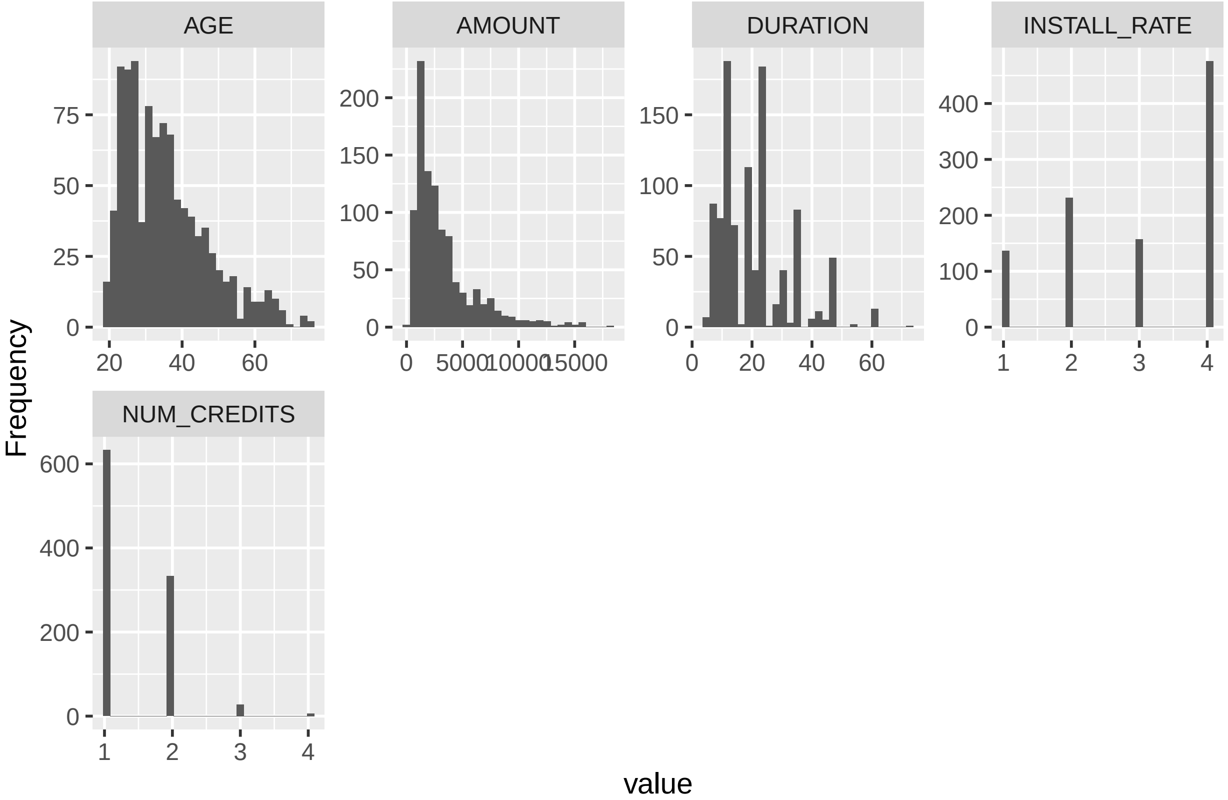


The plot helps us to see the percentage of continuous variable, the percentage of discrete variables and whether or not some observations are missing.

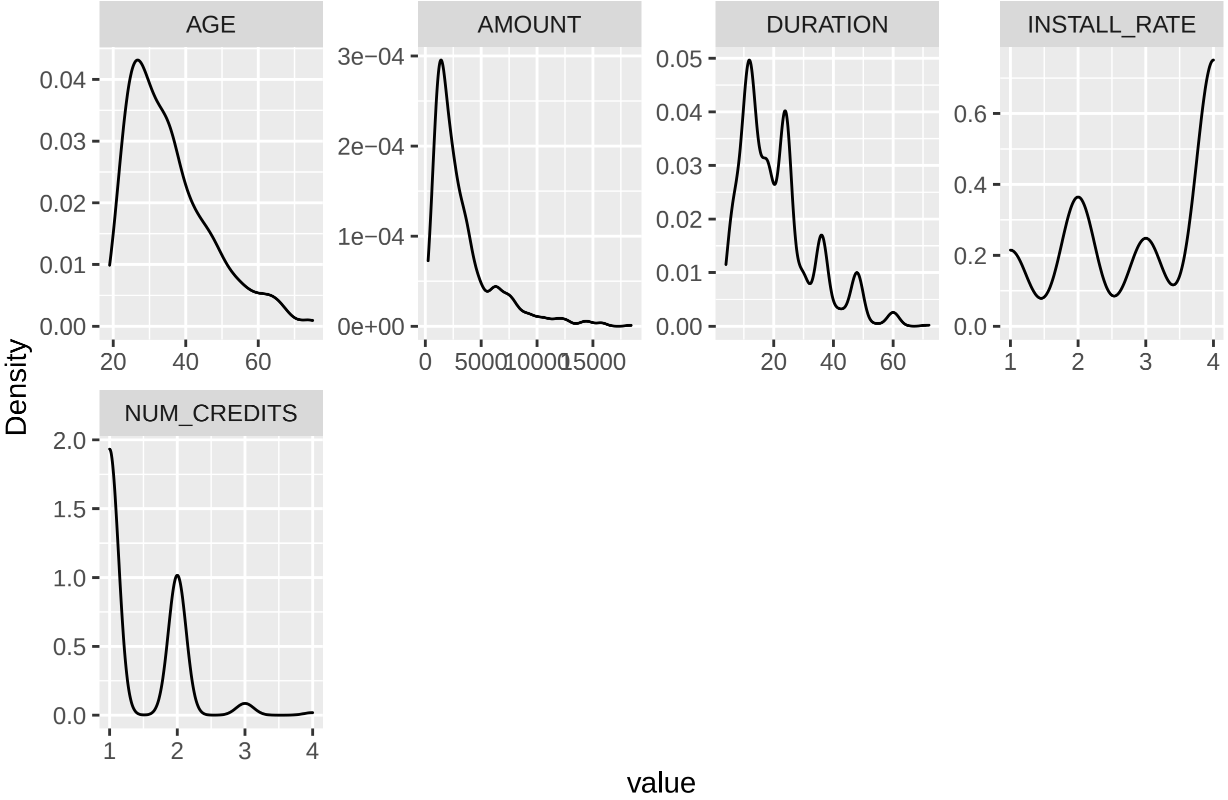
Visualization of the data

First, we plot all the continuous variables into histograms and their corresponding density plots.

```
plot_histogram(German_credit)
```



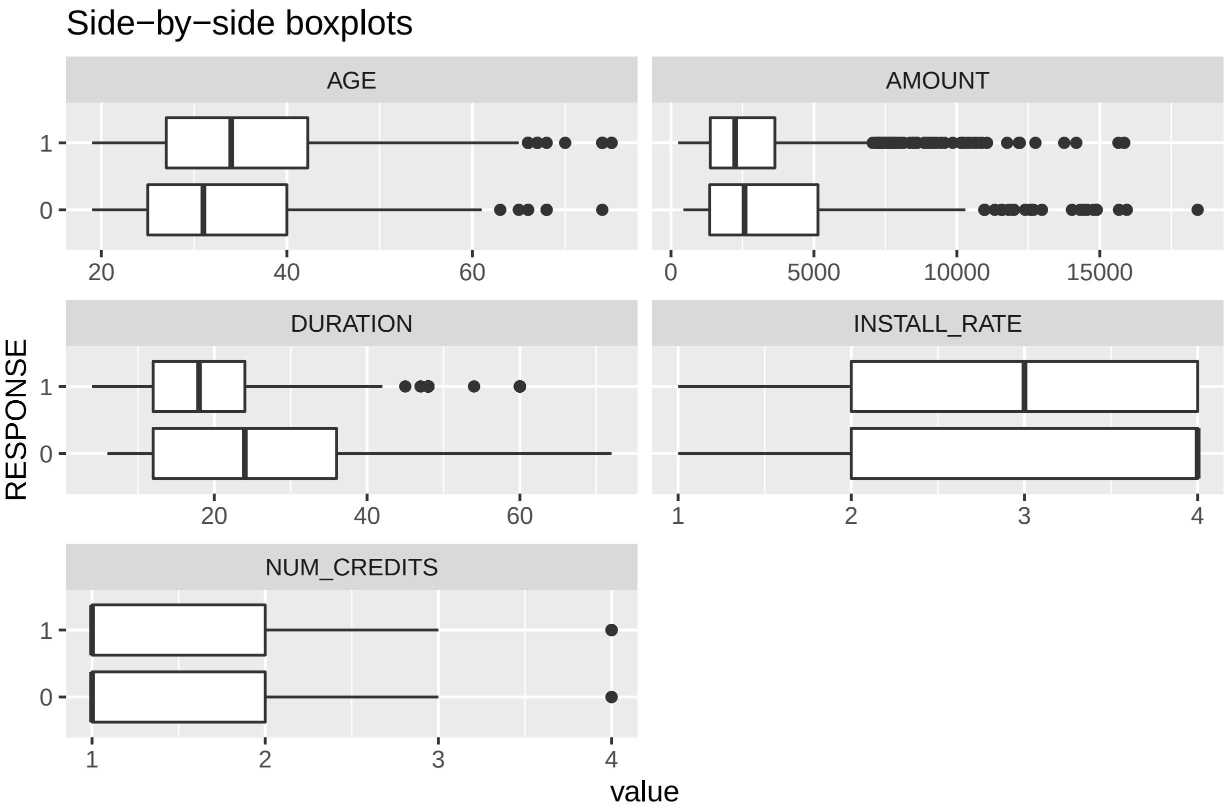
```
plot_density(German_credit)
```



Our first notice is that the data are not really normally distributed. Some of them are right-tailed.

We can look at the tails and outliers more carefully through boxplots.

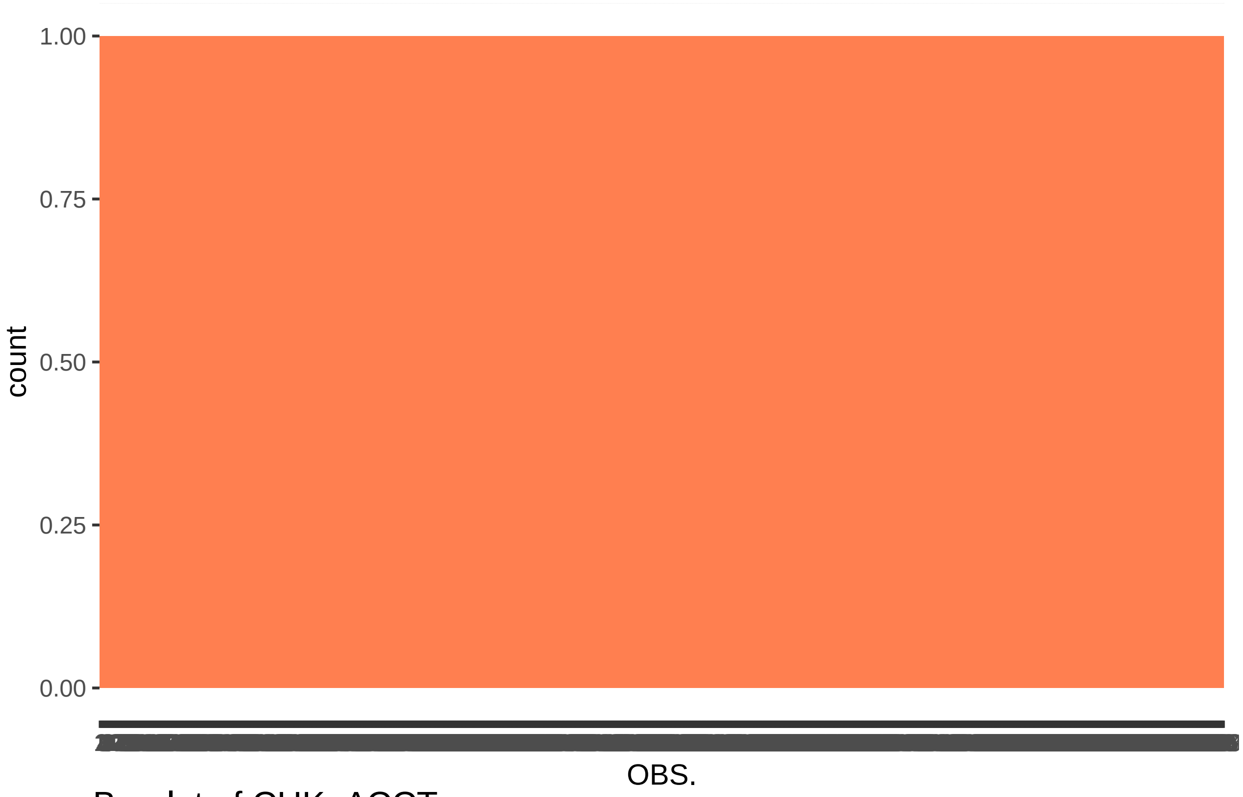
```
plot_boxplot(German_credit, by = 'RESPONSE',  ncol = 2,
             title = "Side-by-side boxplots")
```

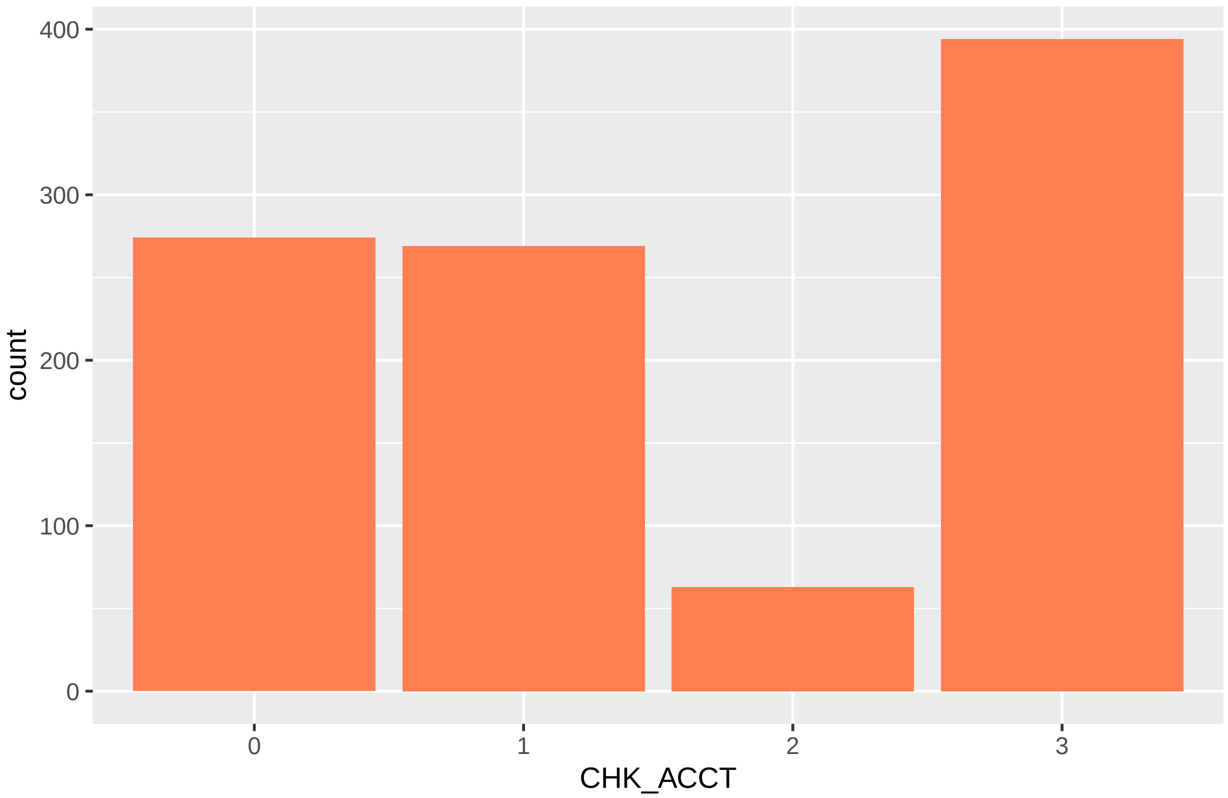


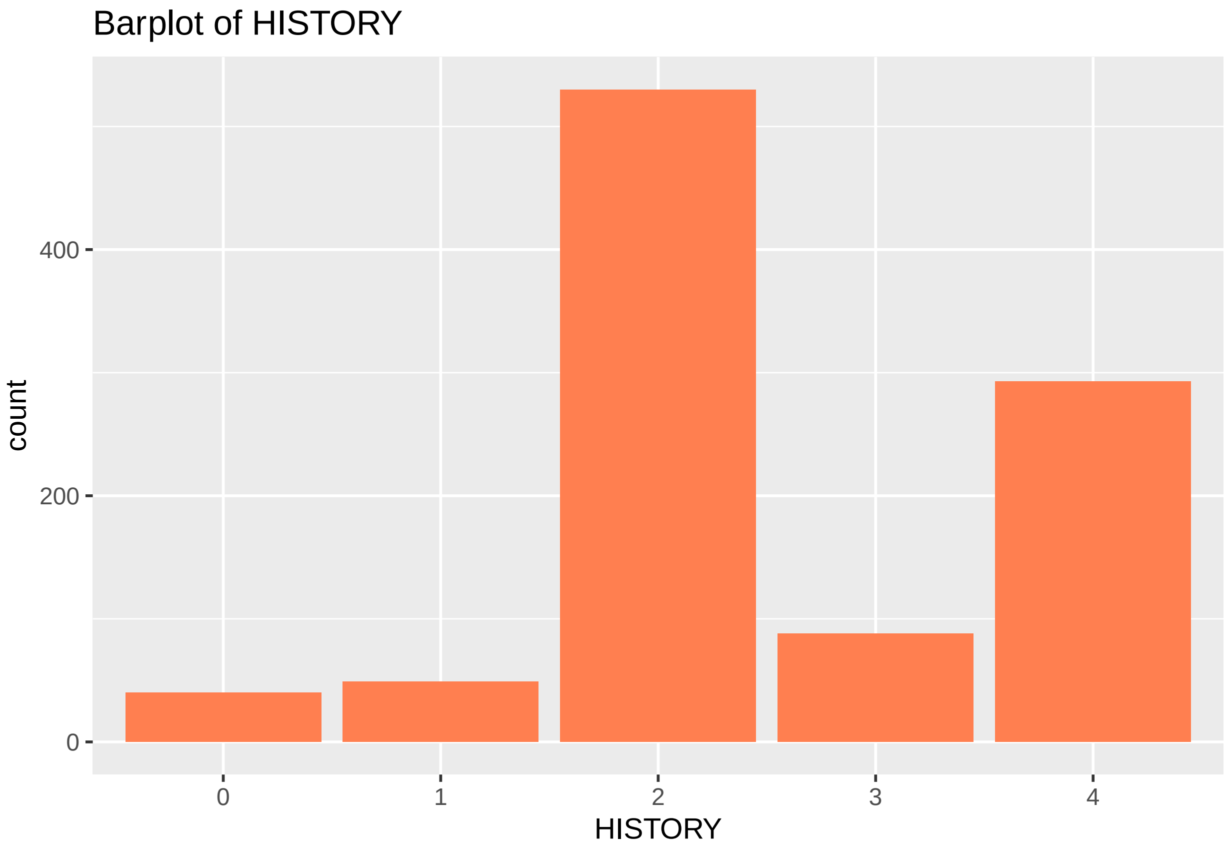
This seems not to be disturbing. It makes sense that only a few people has a big credit amount. However it seems that the 'bad' clients tends to ask for bigger credit amount than 'good' clients.

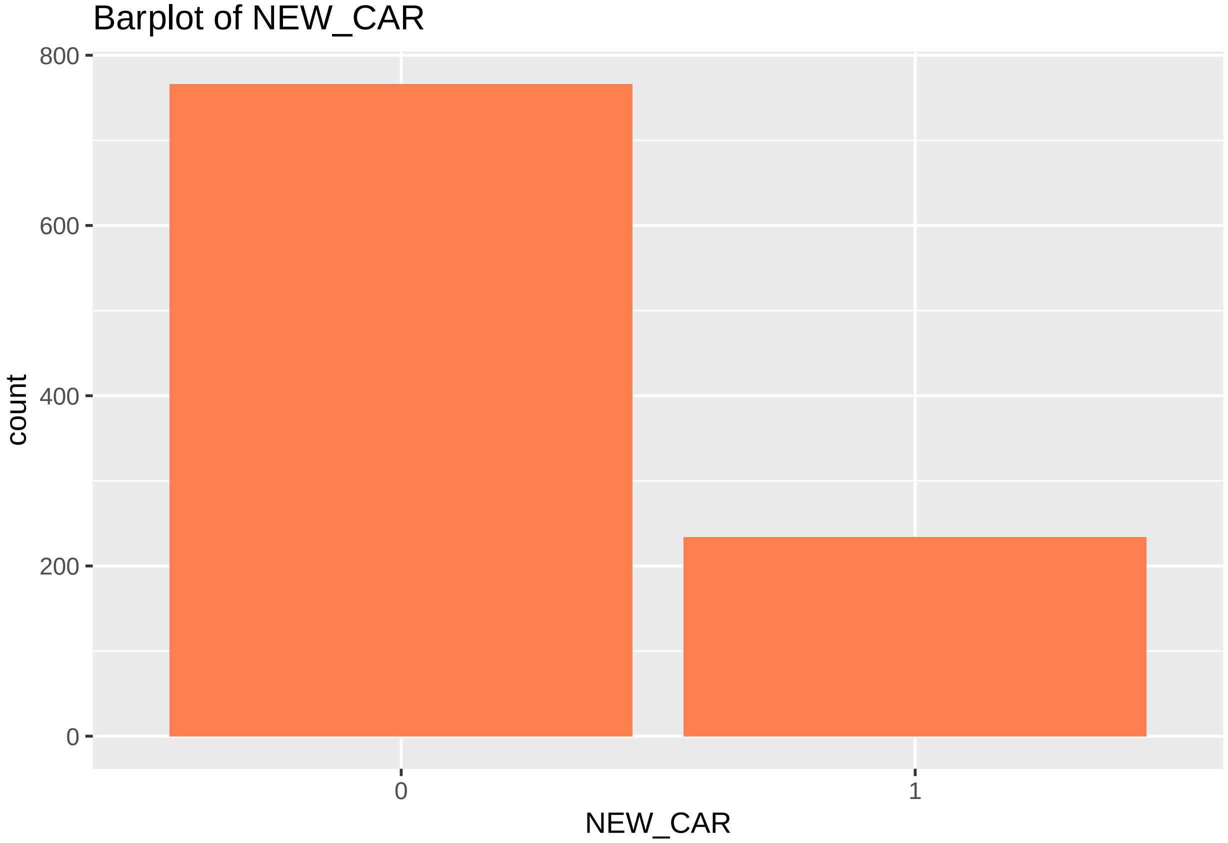
Now, we can make some barplots of the categorical variables.

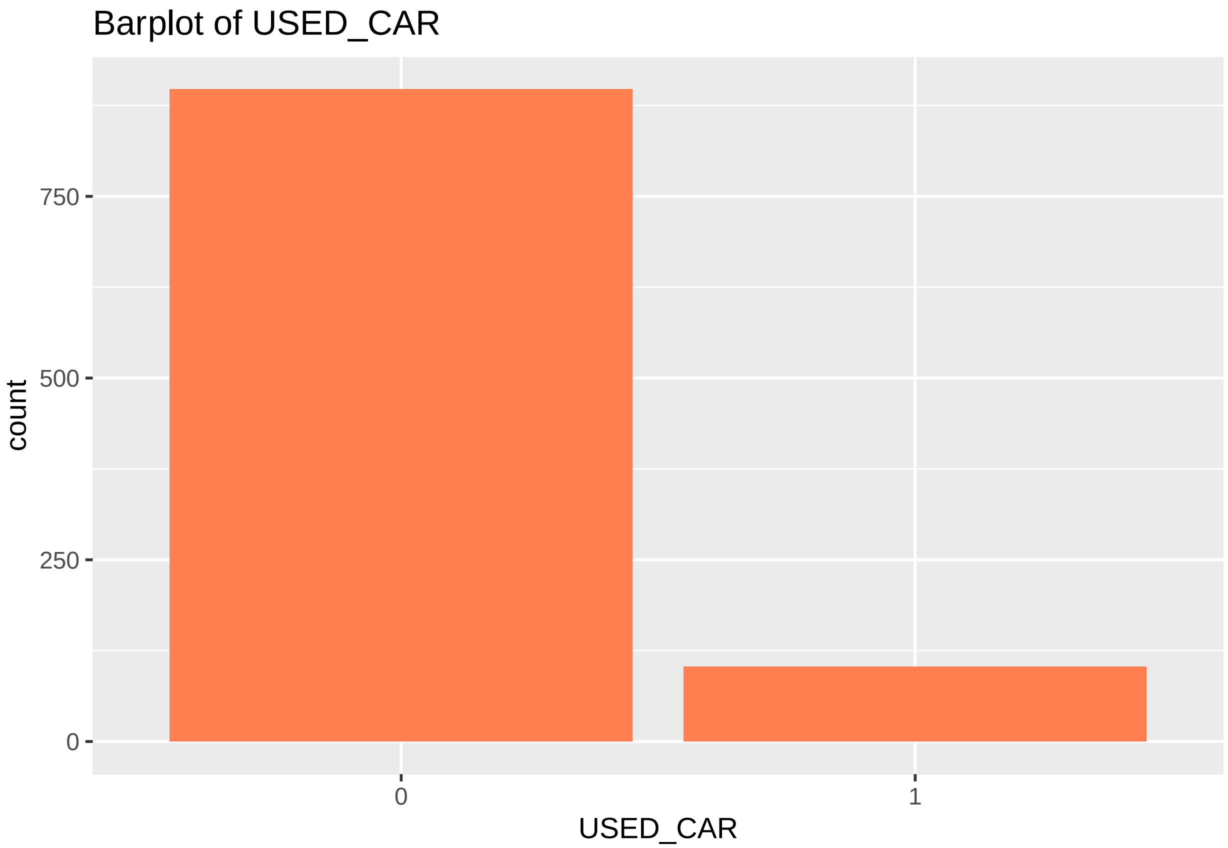
```
for (i in 1:ncol(German_credit)){
  if (class(German_credit[,i])=="factor"){
    print(ggplot(German_credit) +
            geom_bar(aes(x=German_credit[,i]), fill = "coral") +
            ggtitle(paste("Barplot of", colnames(German_credit)[i])) +
            labs(x = paste(colnames(German_credit)[i])))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

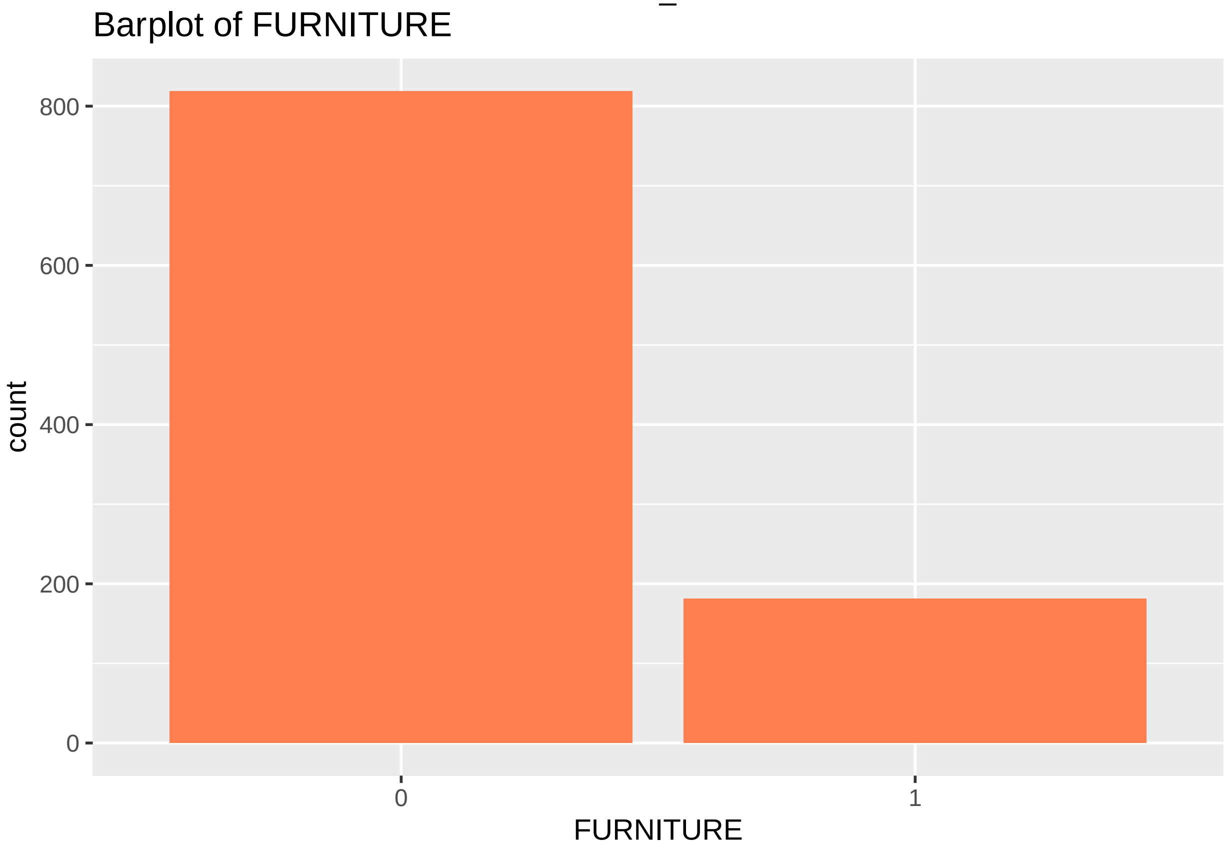

Barplot of OBS.

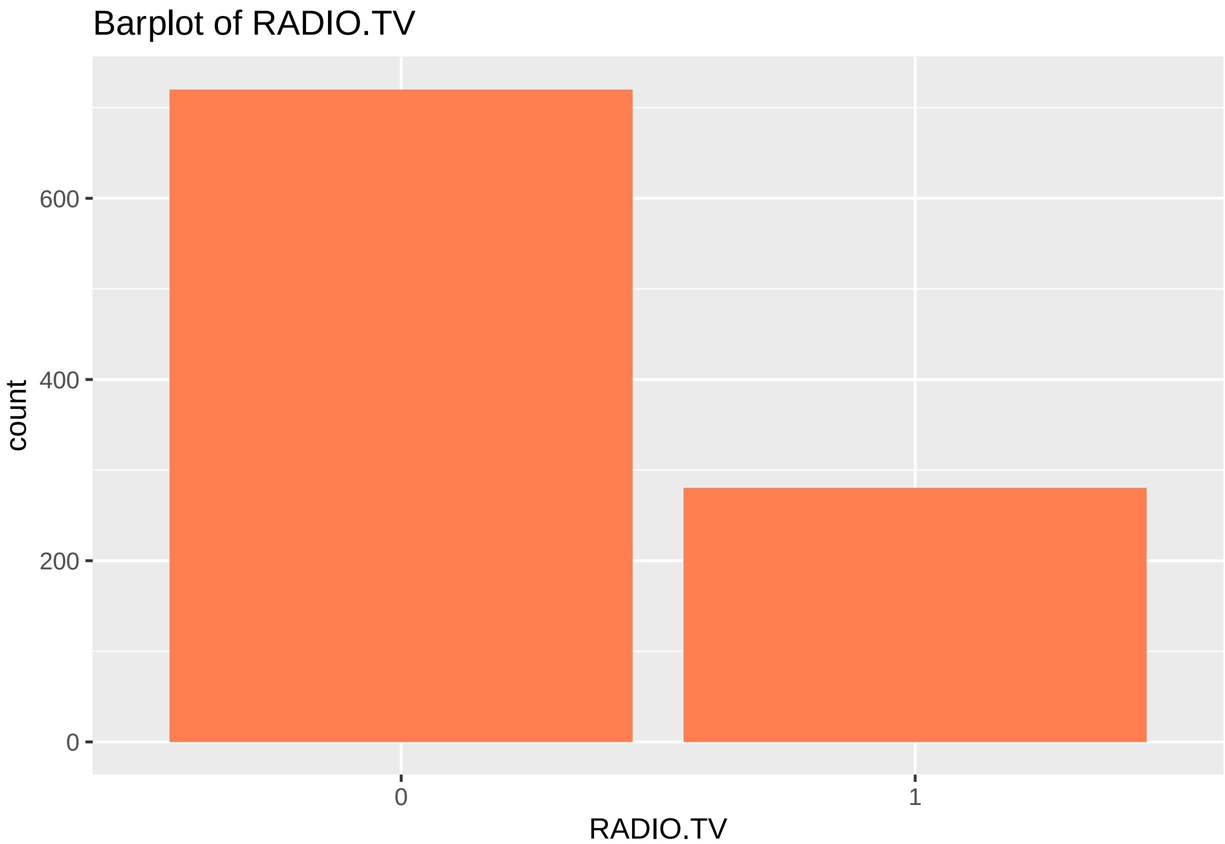

Barplot of CHK_ACCT

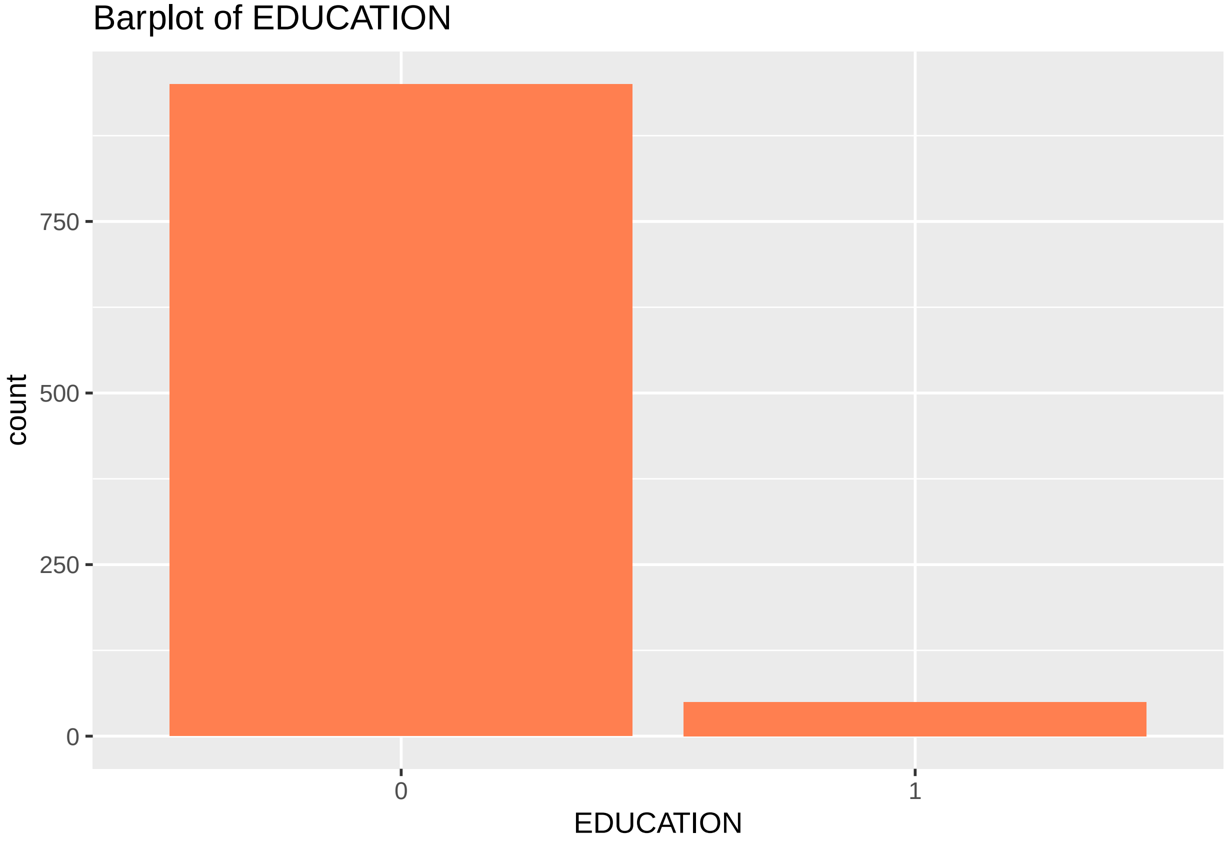

Barplot of HISTORY
count
400
200
0
0
1
2
3
4
HISTORY

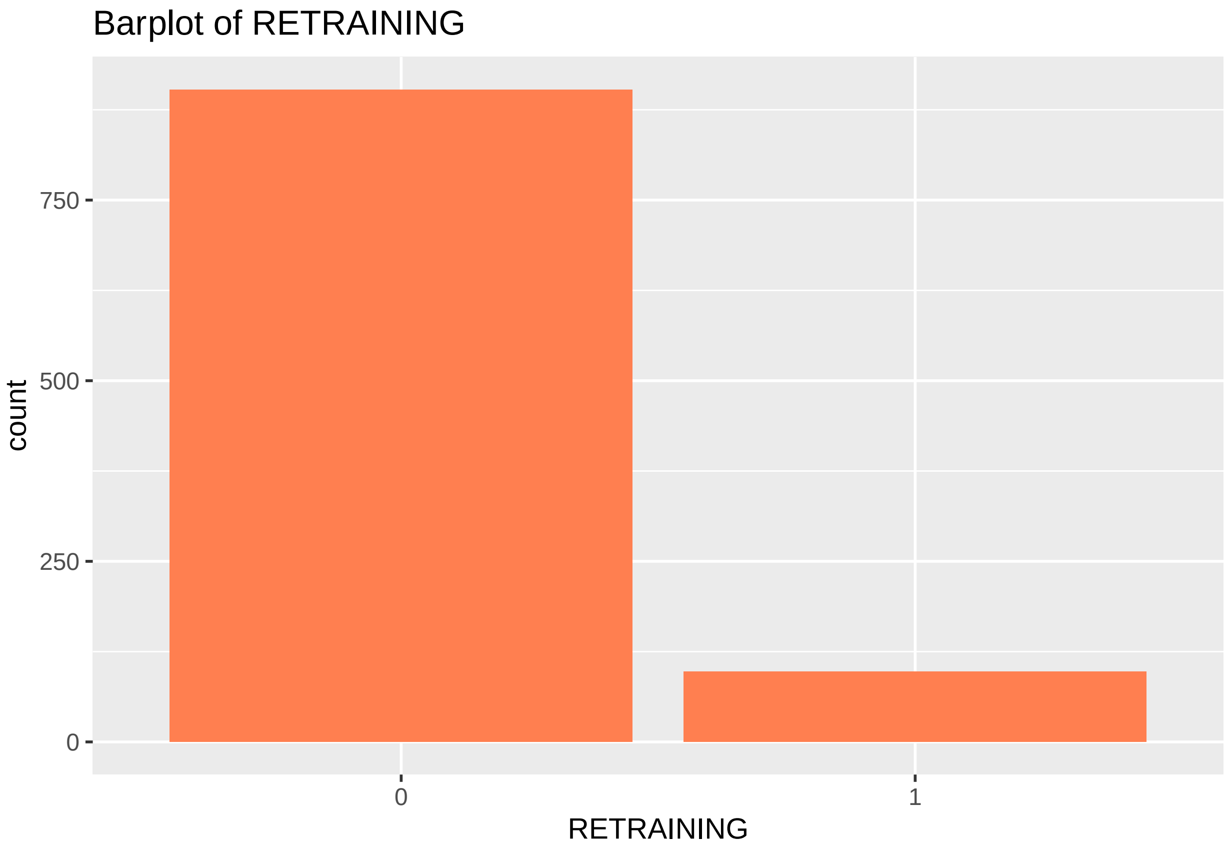

Barplot of NEW_CAR
count
800
600
400
200
0
0
1
NEW_CAR

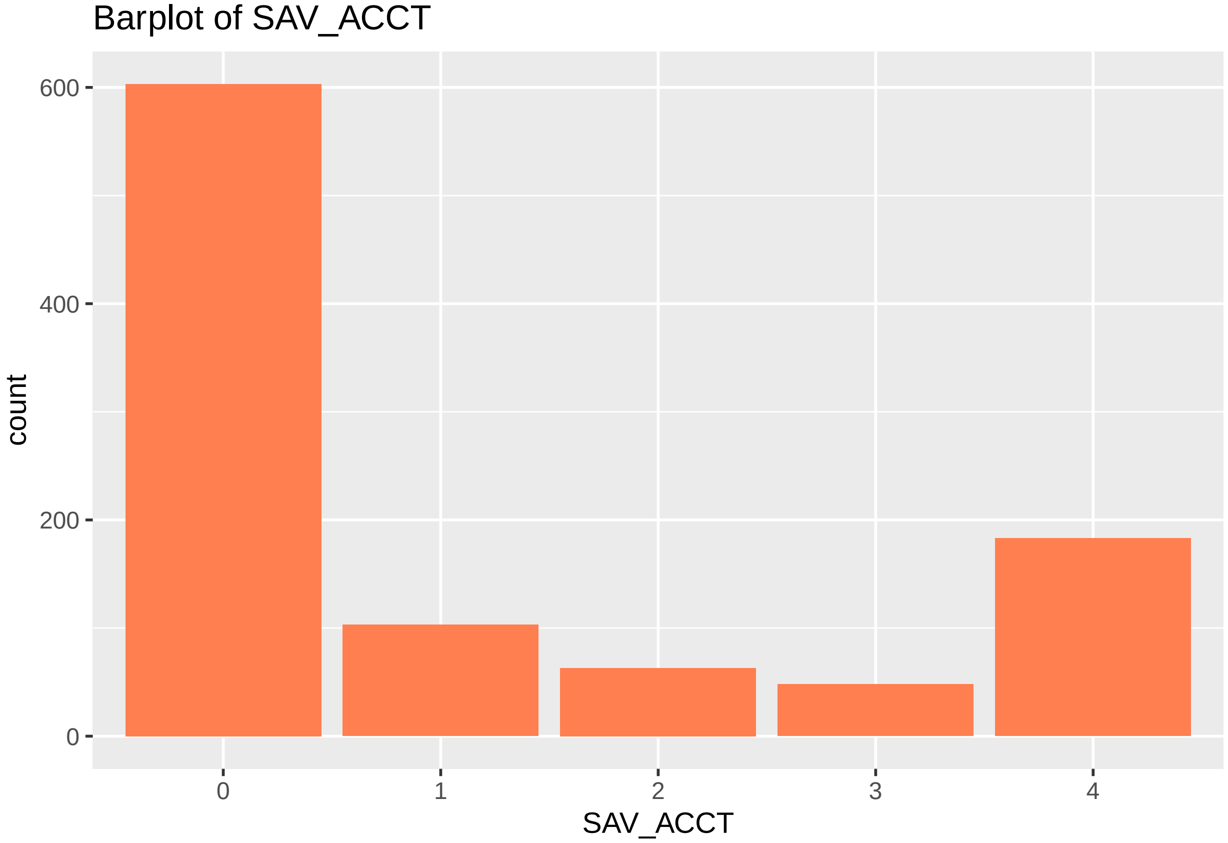

Barplot of USED_CAR
count
750
500
250
0
0
1
USED_CAR

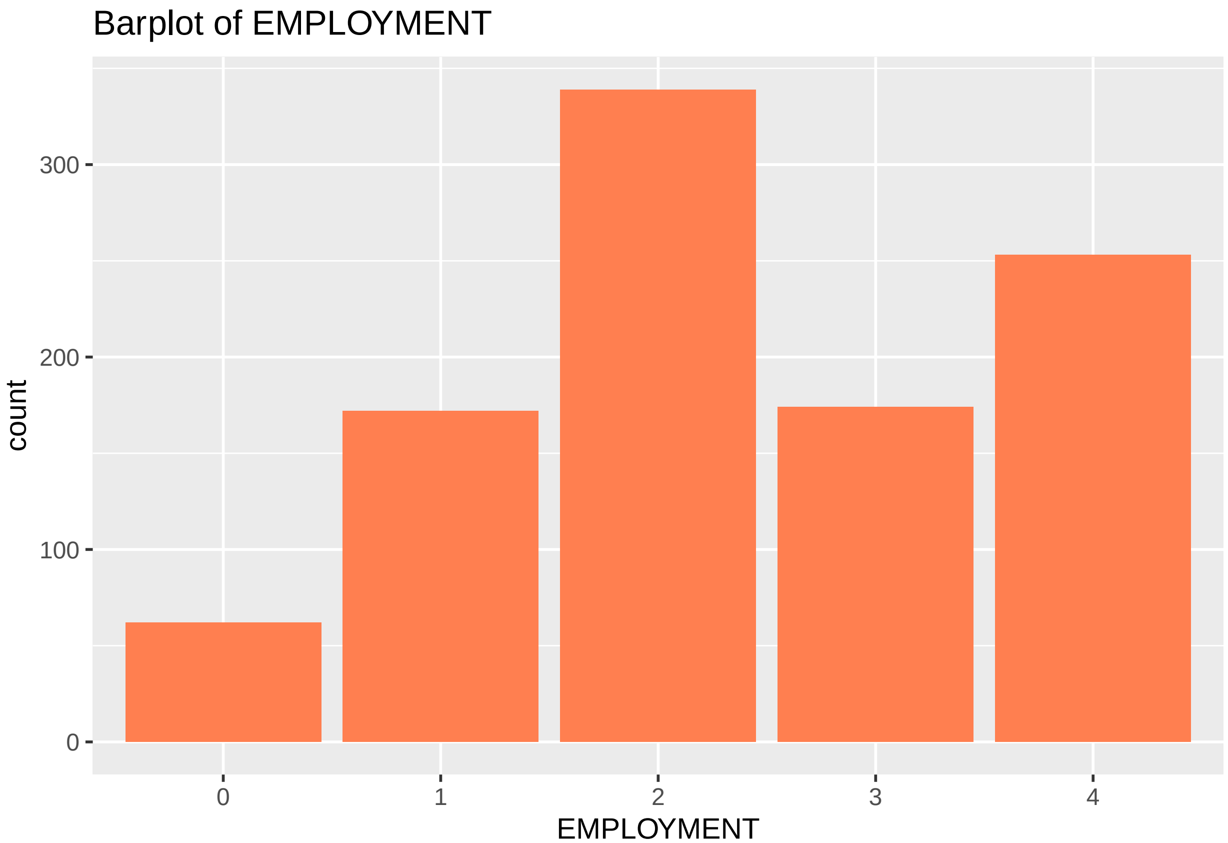

Barplot of FURNITURE
count
800
600
400
200
0
0
1
FURNITURE

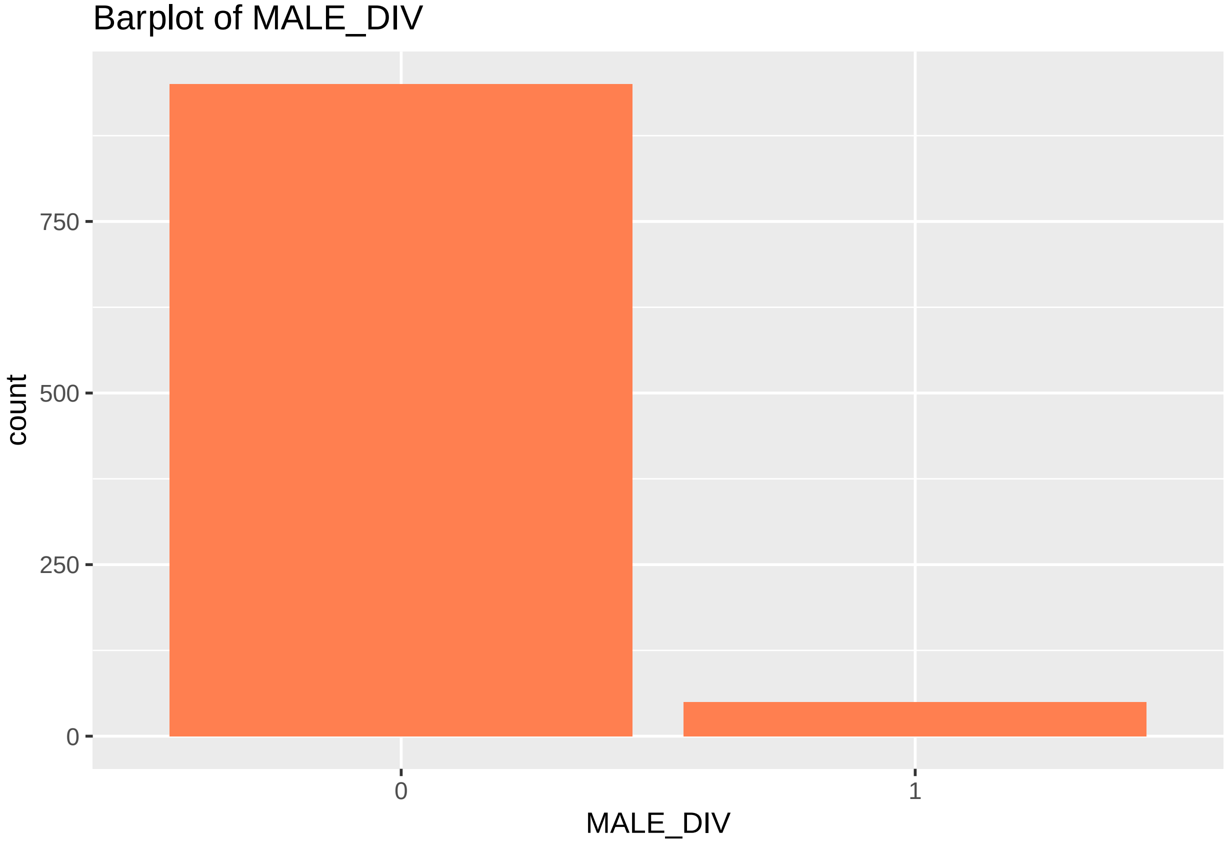

Barplot of RADIO.TV
count
600
400
200
0
0
1
RADIO.TV

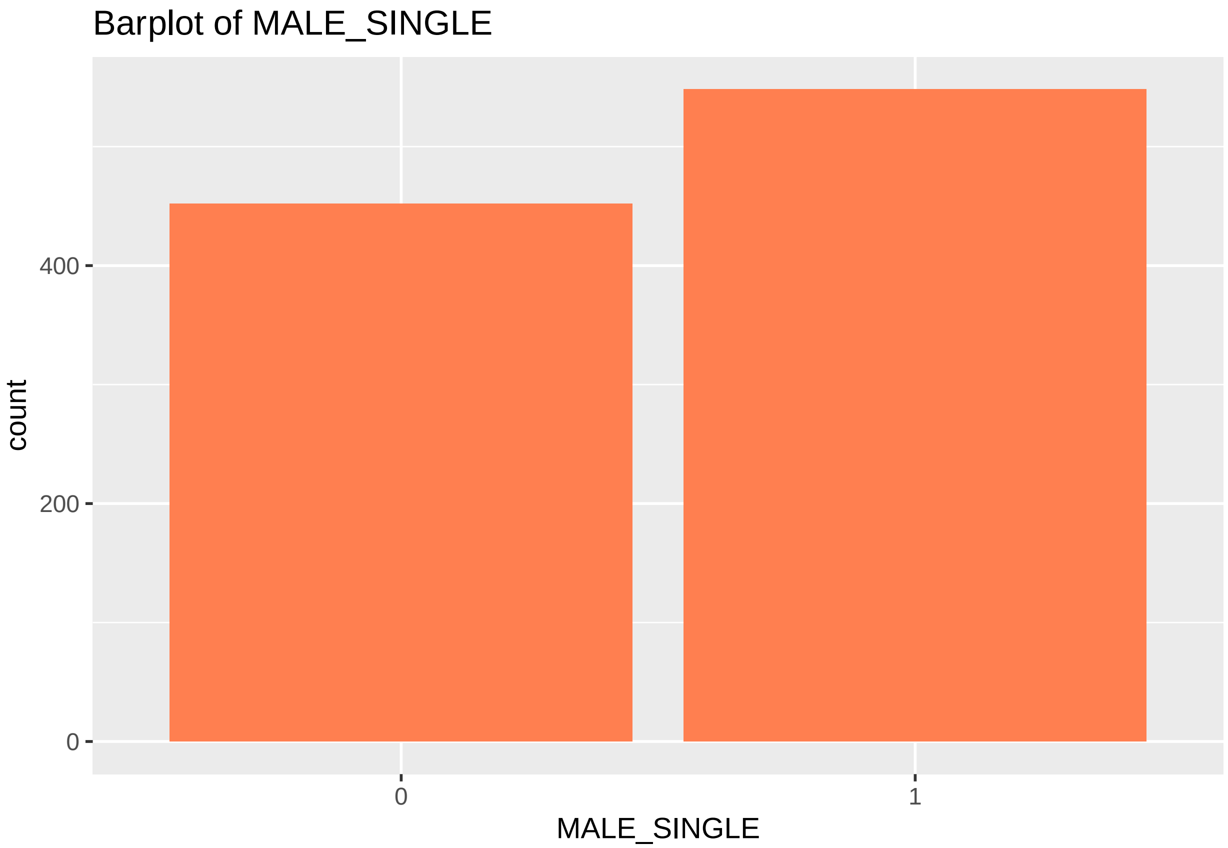

Barplot of EDUCATION
count
750
500
250
0
0
1
EDUCATION

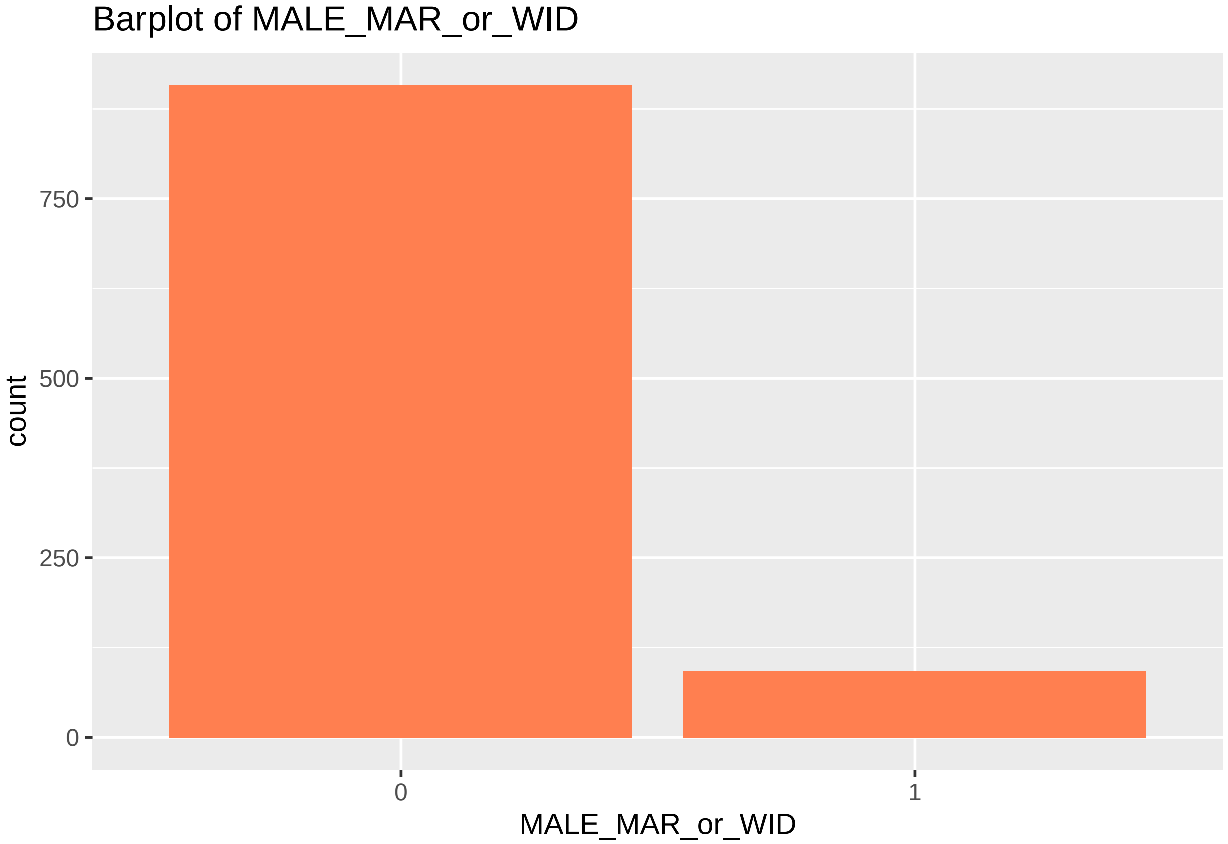

Barplot of RETRAINING
count
750
500
250
0
0
1
RETRAINING

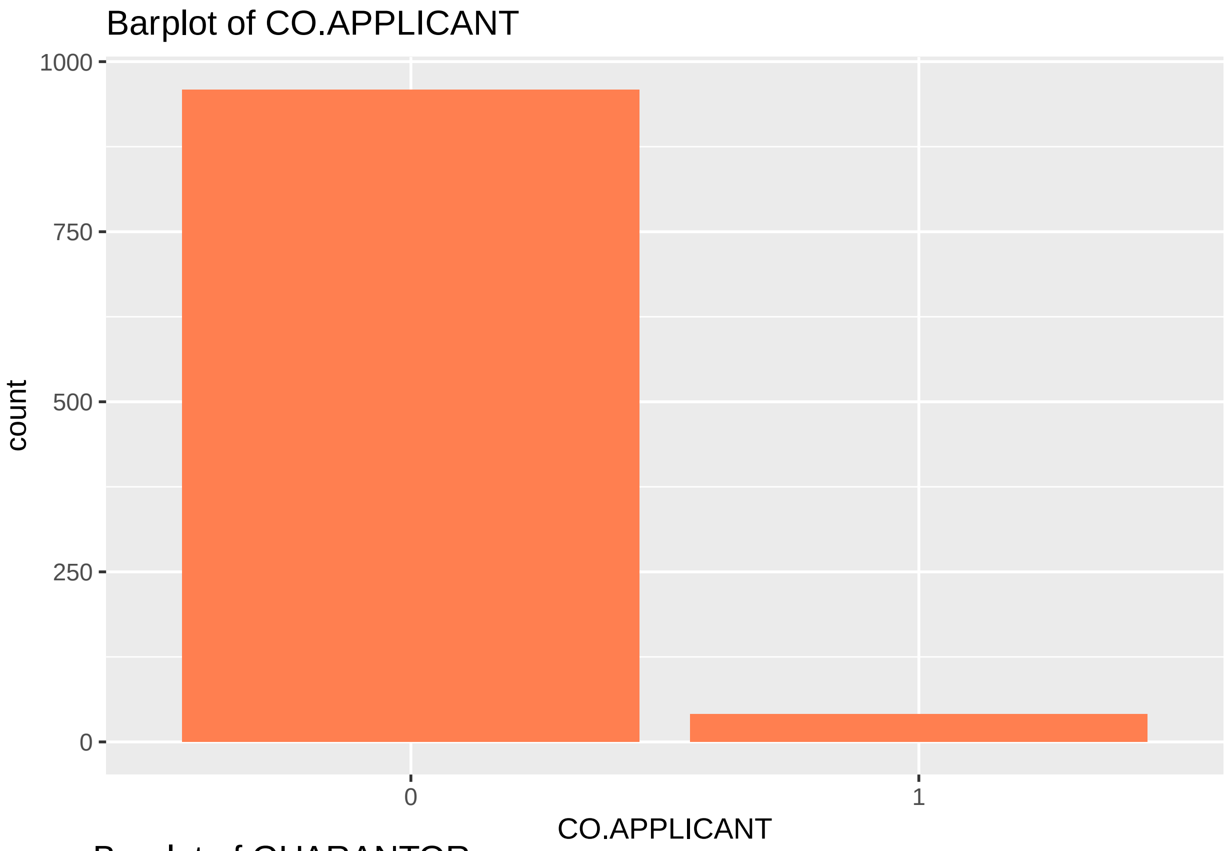

Barplot of SAV_ACCT
count
600
400
200
0
0
1
2
3
4
SAV_ACCT

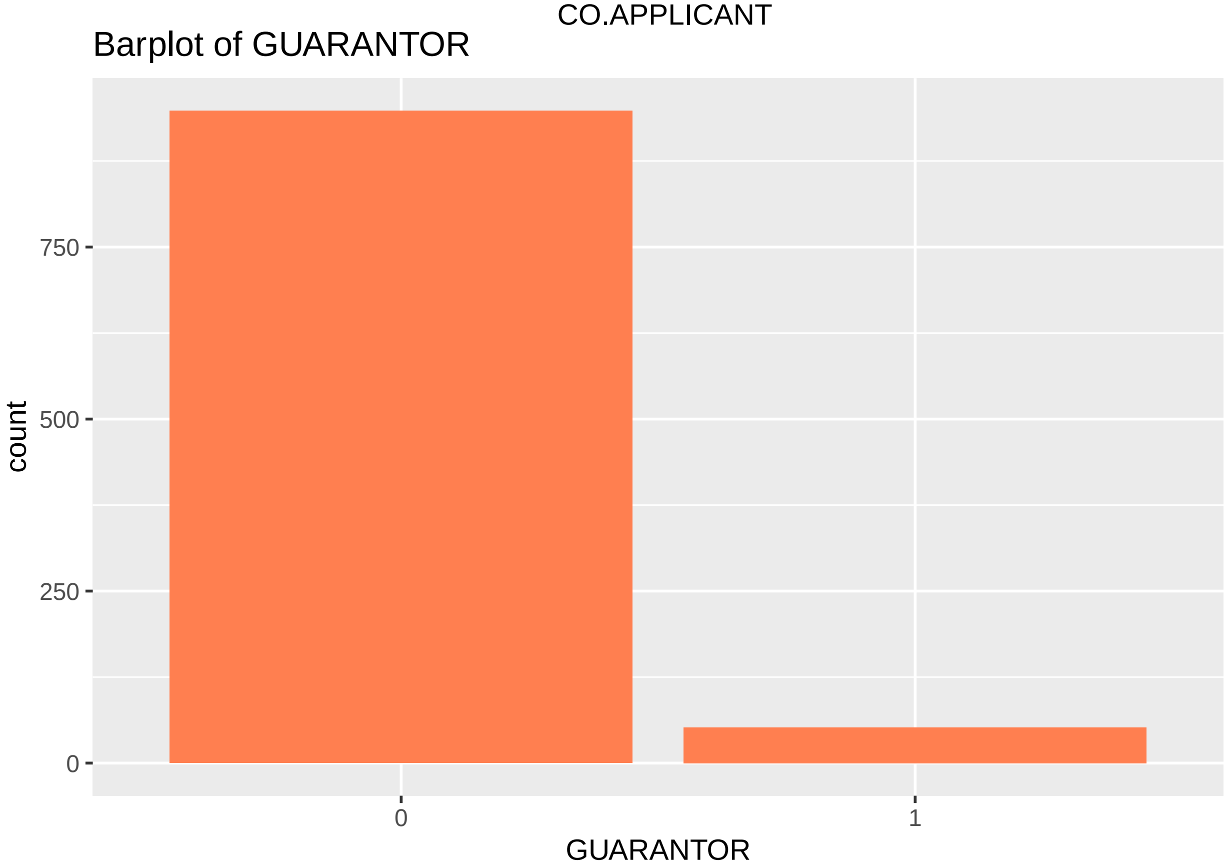

Barplot of EMPLOYMENT
count
300
200
100
0
0
1
2
3
4
EMPLOYMENT

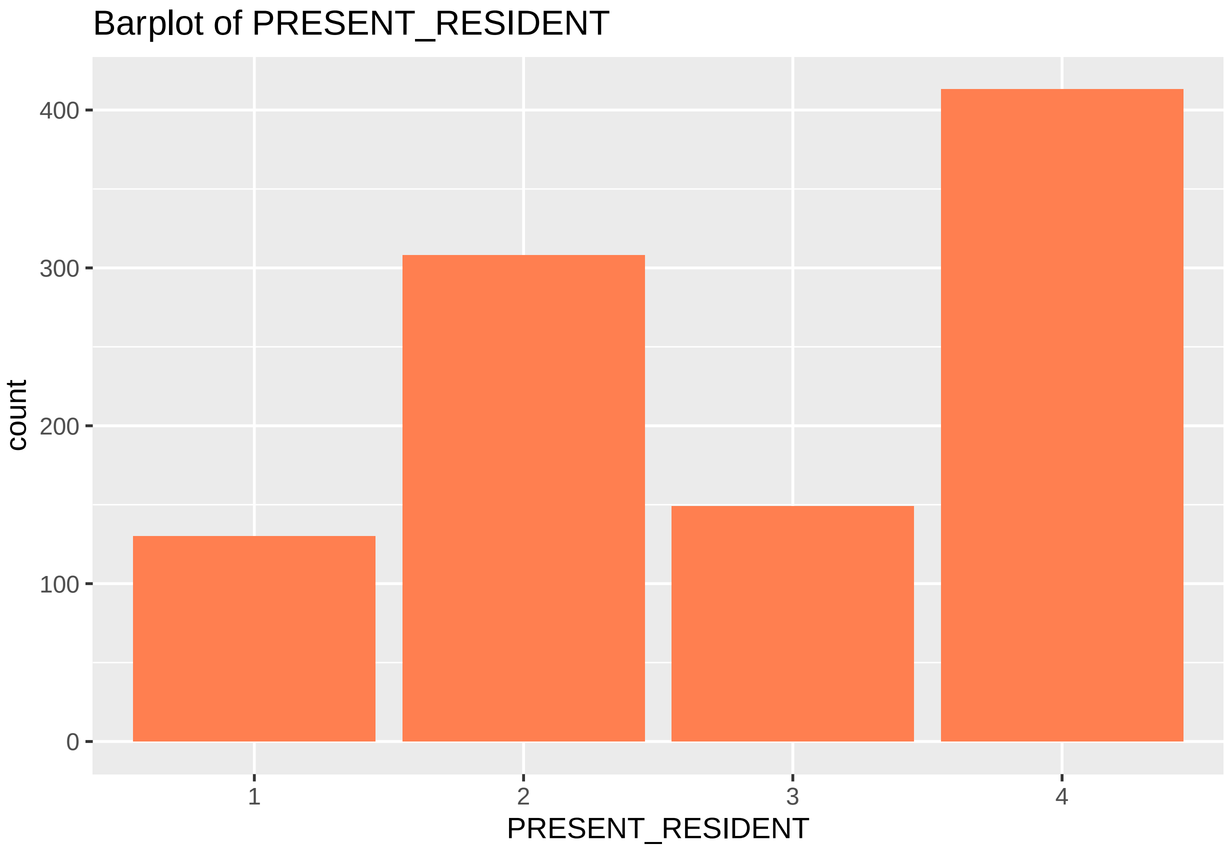

Barplot of MALE_DIV
count
750
500
250
0
0
1
MALE_DIV

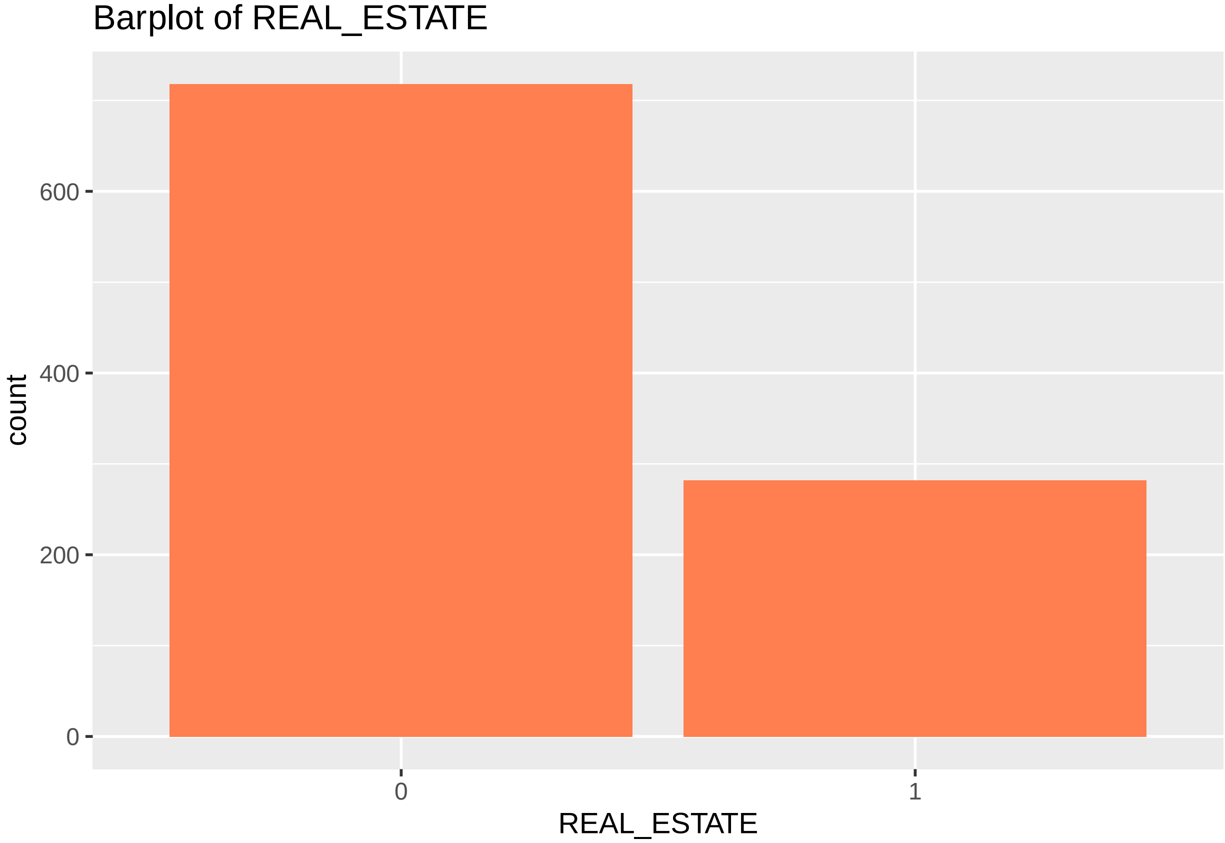

Barplot of MALE_SINGLE
count
400
200
0
0
1
MALE_SINGLE


Barplot of MALE_MAR_or_WID
count
750
500
250
0
0
1
MALE_MAR_or_WID

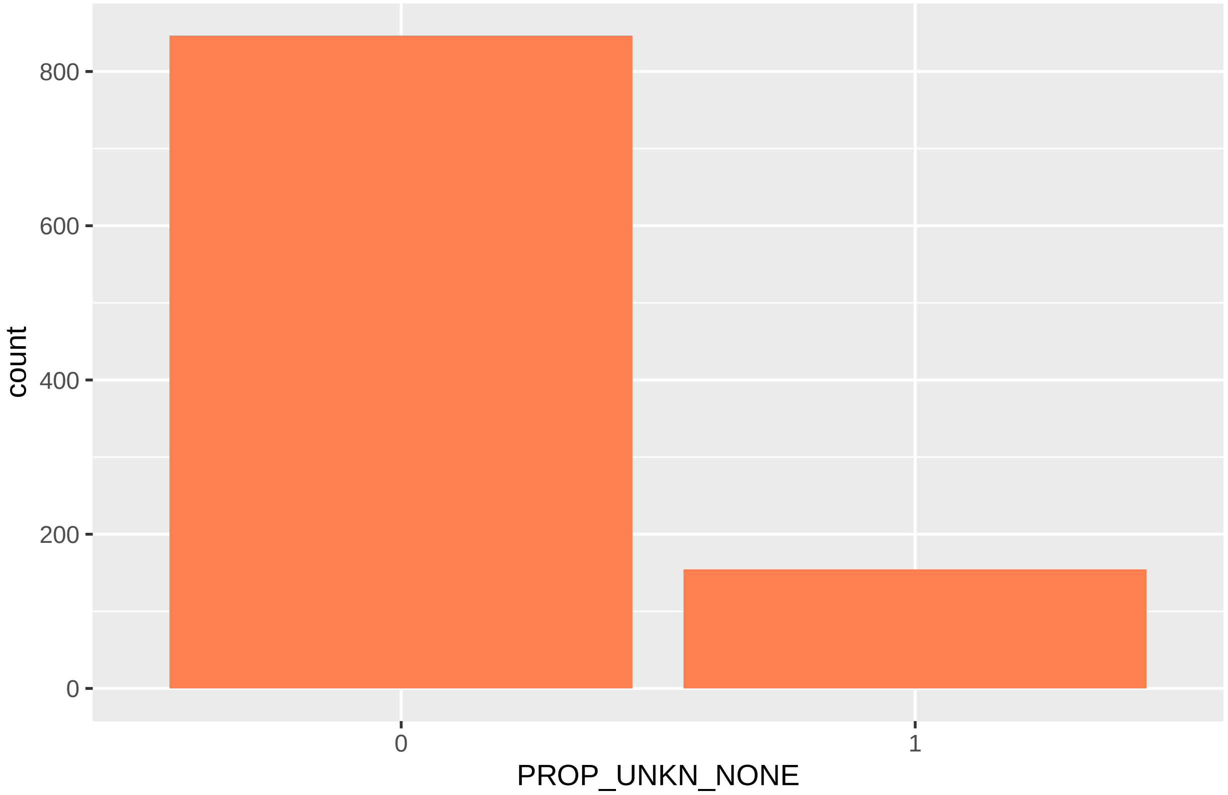

Barplot of CO.APPLICANT
count
1000
750
500
250
0
0
1
CO.APPLICANT


Barplot of GUARANTOR
count
750
500
250
0
0
1
GUARANTOR

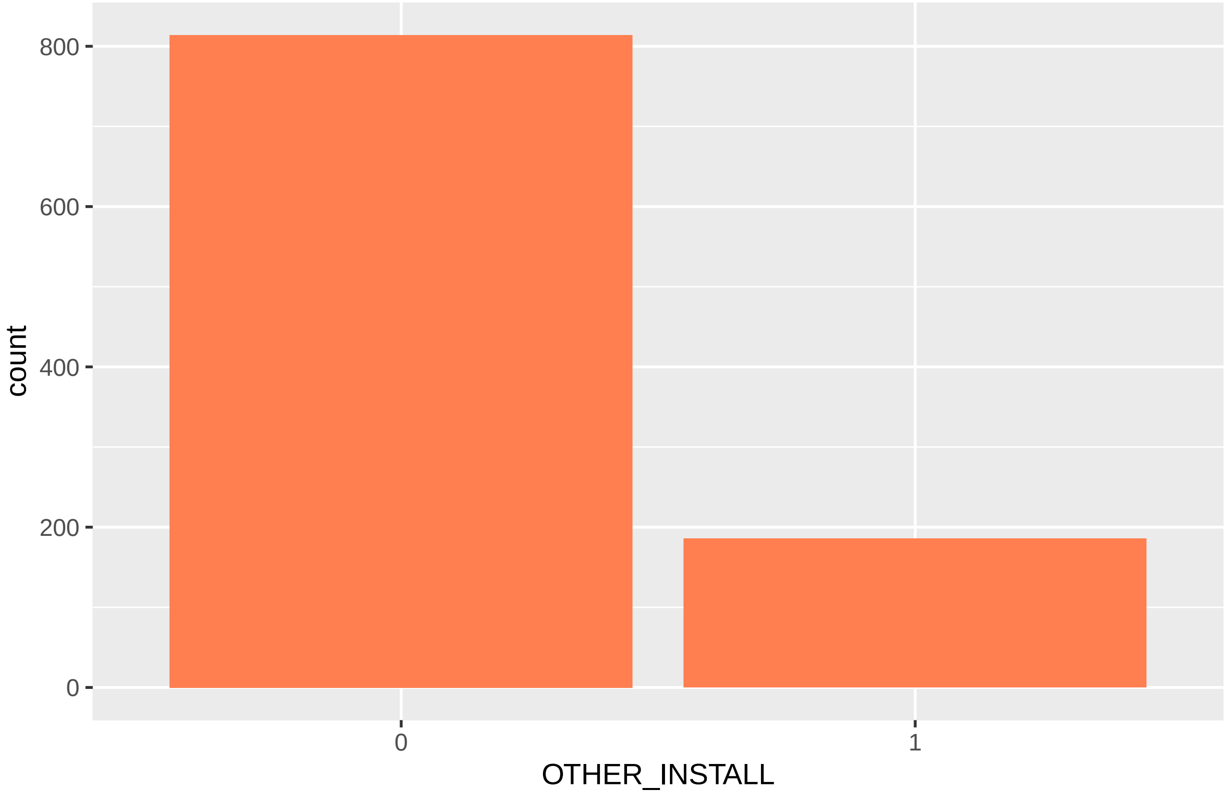

Barplot of PRESENT_RESIDENT
count
400
300
200
100
0
1
2
3
4
PRESENT_RESIDENT


Barplot of REAL_ESTATE
count
600
400
200
0
0
1
REAL_ESTATE

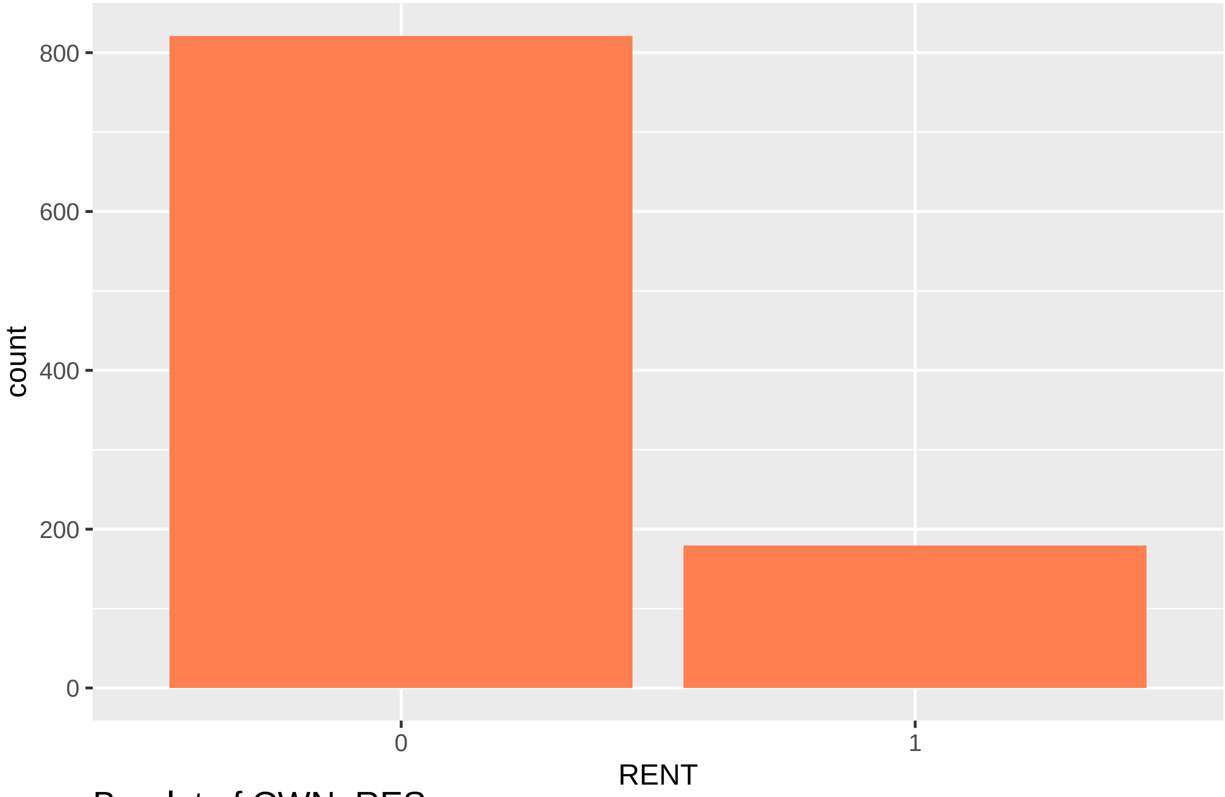
Barplot of PROP_UNKN_NONE



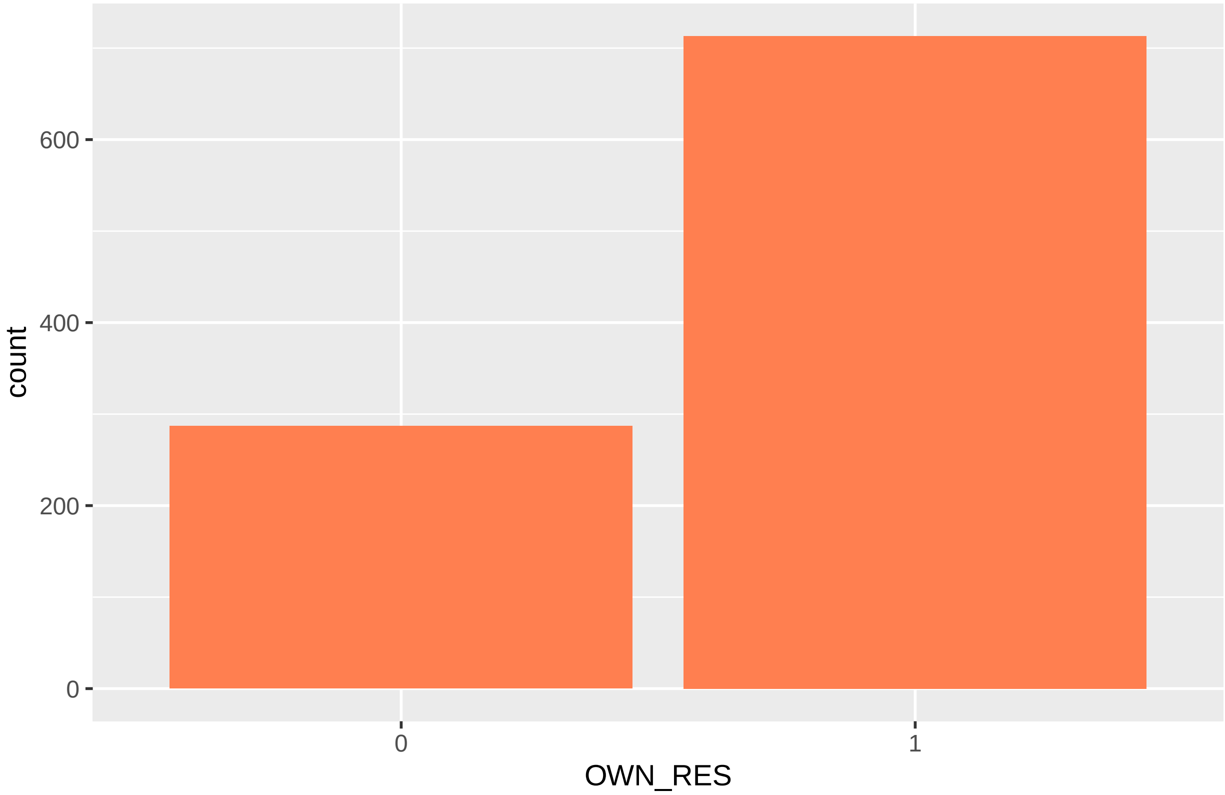
Barplot of OTHER_INSTALL



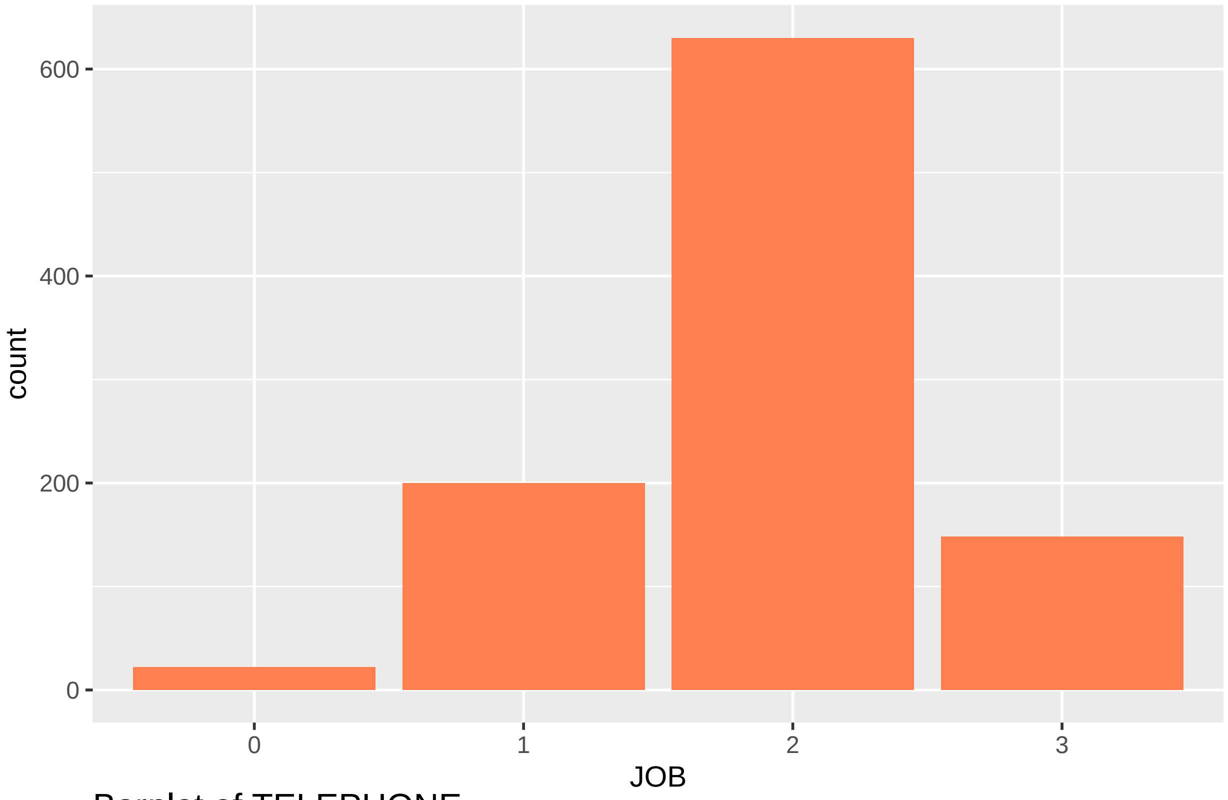
Barplot of RENT



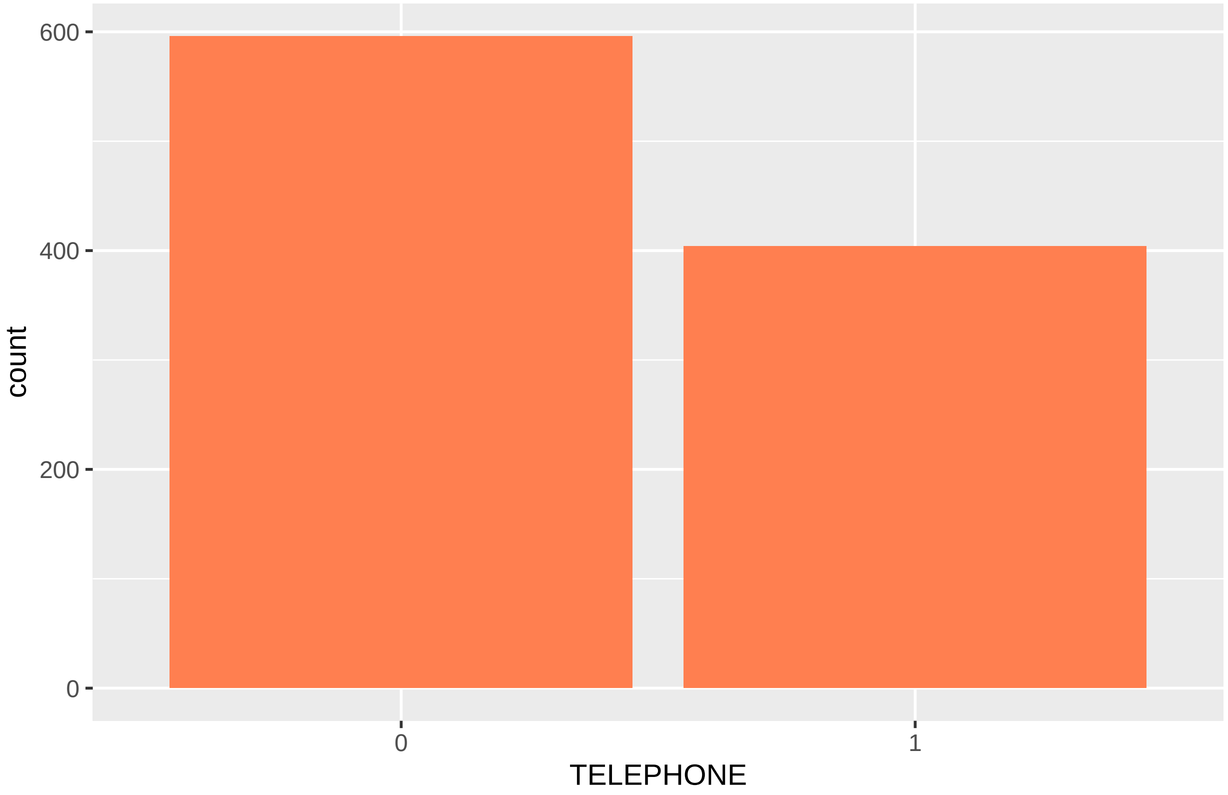
Barplot of OWN_RES

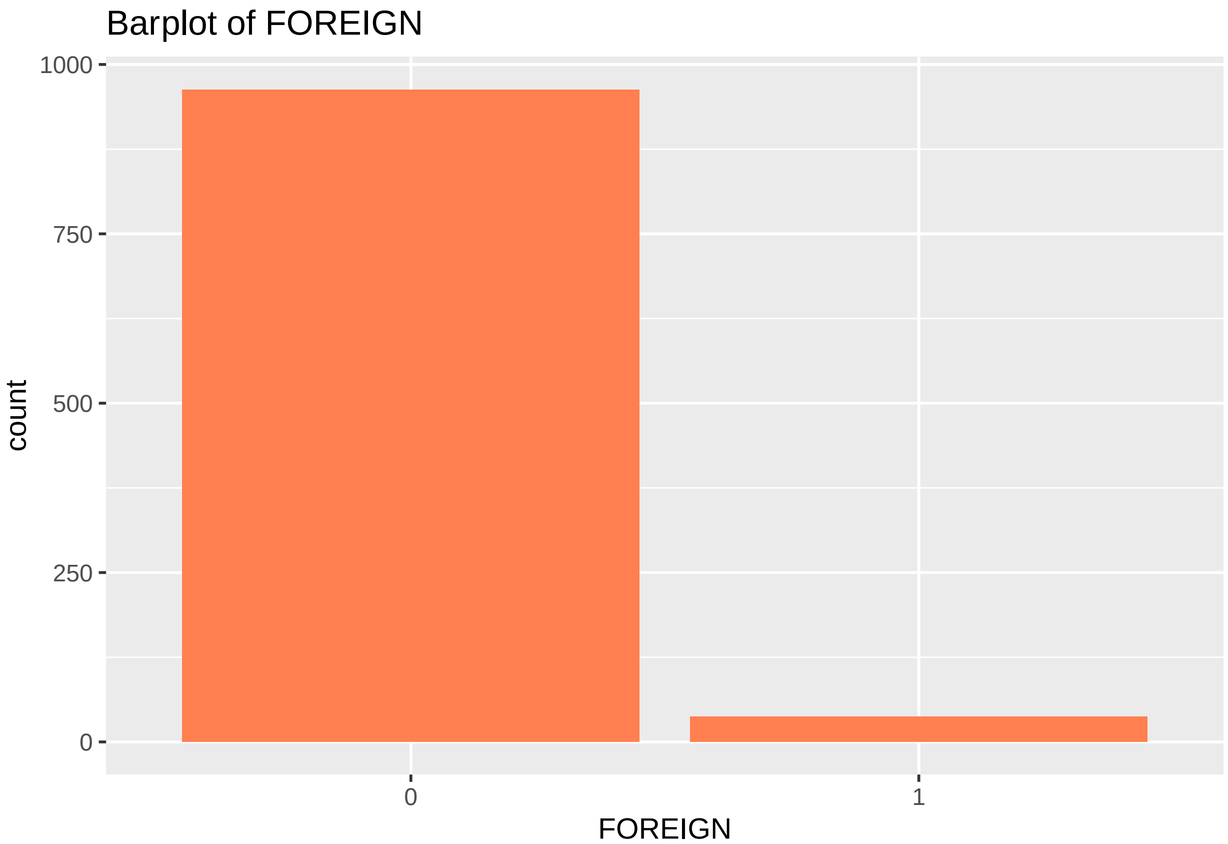


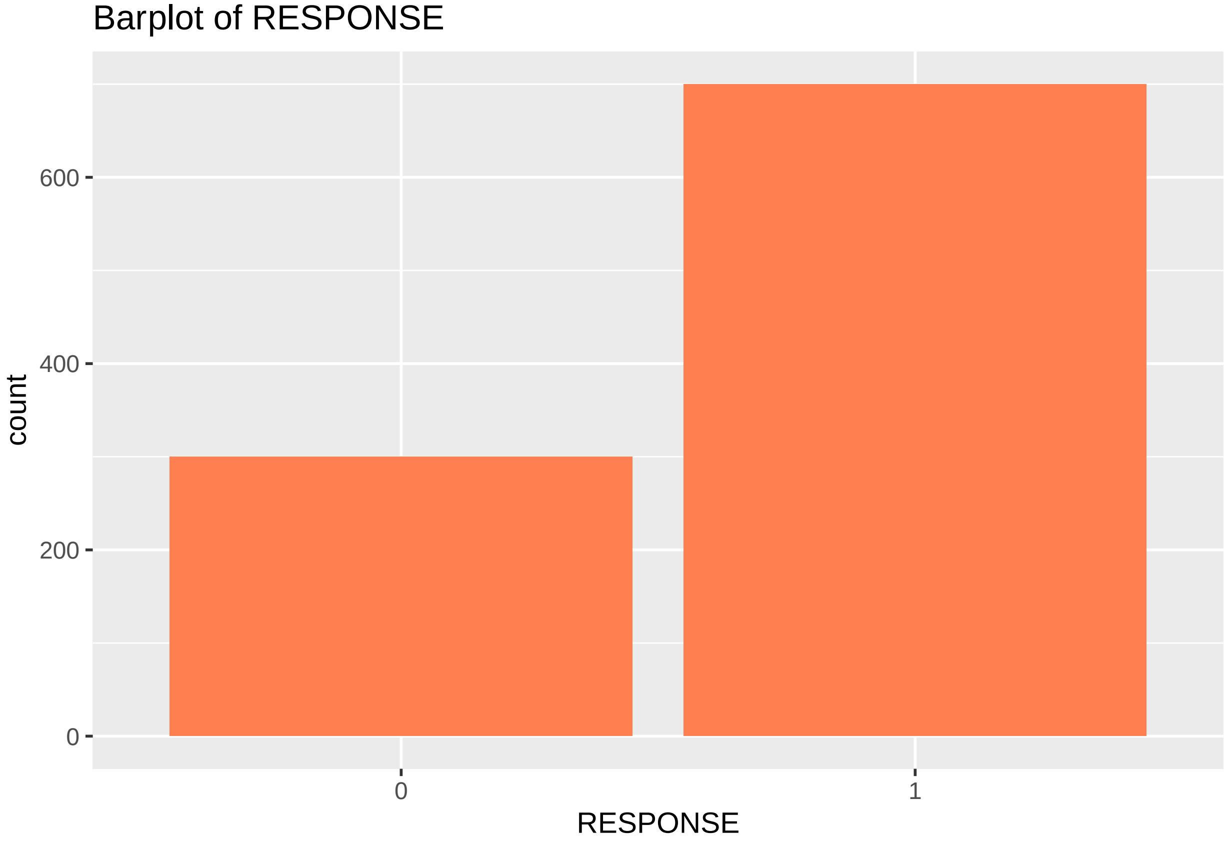
Barplot of JOB



Barplot of 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)
- We have more information on people that are 'good' applicants and less information on 'bad' applicants. It would be better to have more information on 'bad' applicants as well in order to make good predictions with models. We will have to take this into account later. (RESPONSE)

A general summary can be done.

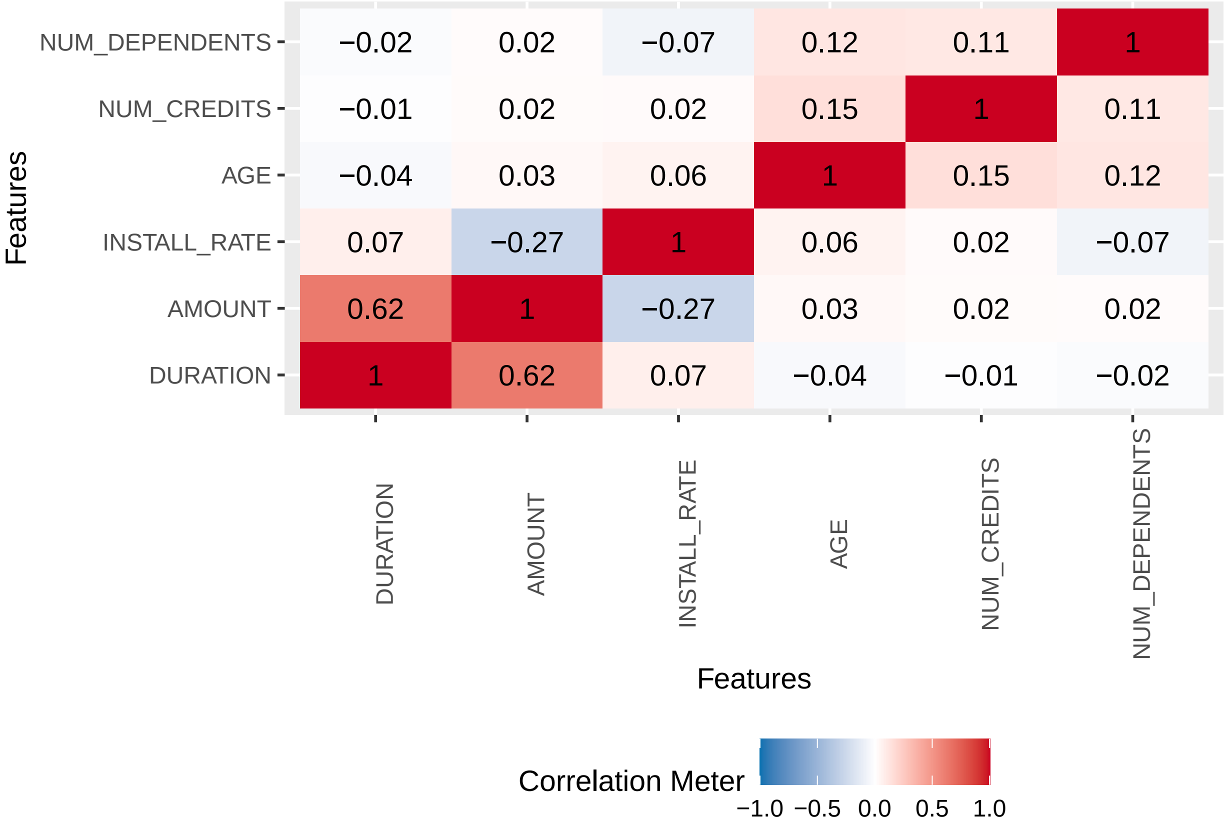
```
dfSummary(German_credit, style = 'grid')
```

```
## Data Frame Summary
## German_credit
## Dimensions: 1000 x 32
## Duplicates: 0
##
## +-----+------------------+-----------------------------+---------------------+-----------------------
## | No  | Variable         | Stats / Values              | Freqs (% of Valid)  | Graph
## +=====+==================+=============================+=====================+=======================
## | 1   | OBS.             | 1. 1                        |   1 ( 0.1%)         |
## |     | [factor]         | 2. 2                        |   1 ( 0.1%)         |
## |     |                  | 3. 3                        |   1 ( 0.1%)         |
## |     |                  | 4. 4                        |   1 ( 0.1%)         |
## |     |                  | 5. 5                        |   1 ( 0.1%)         |
## |     |                  | 6. 6                        |   1 ( 0.1%)         |
## |     |                  | 7. 7                        |   1 ( 0.1%)         |
## |     |                  | 8. 8                        |   1 ( 0.1%)         |
## |     |                  | 9. 9                        |   1 ( 0.1%)         |
## |     |                  | 10. 10                      |   1 ( 0.1%)         |
## |     |                  | [ 990 others ]              | 990 (99.0%)         | IIIIIIIIIIIIIIIIIII
## +-----+------------------+-----------------------------+---------------------+-----------------------
## | 2   | CHK_ACCT         | 1. 0                        | 274 (27.4%)         | IIIII
## |     | [factor]         | 2. 1                        | 269 (26.9%)         | IIIII
## |     |                  | 3. 2                        |  63 ( 6.3%)         | I
## |     |                  | 4. 3                        | 394 (39.4%)         | IIIIIII
## +-----+------------------+-----------------------------+---------------------+-----------------------
## | 3   | DURATION         | Mean (sd) : 20.9 (12.1)     | 33 distinct values  |     :
## |     | [numeric]        | min < med < max:            |                     |   : :
## |     |                  | 4 < 18 < 72                 |                     | . : :
## |     |                  | IQR (CV) : 12 (0.6)         |                     | : : :   .
## |     |                  |                             |                     | : : : : : . :
## +-----+------------------+-----------------------------+---------------------+-----------------------
## | 4   | HISTORY          | 1. 0                        |  40 ( 4.0%)         |
## |     | [factor]         | 2. 1                        |  49 ( 4.9%)         |
## |     |                  | 3. 2                        | 530 (53.0%)         | IIIIIIIIII
## |     |                  | 4. 3                        |  88 ( 8.8%)         | I
## |     |                  | 5. 4                        | 293 (29.3%)         | IIIII
## +-----+------------------+-----------------------------+---------------------+-----------------------
## | 5   | NEW_CAR          | 1. 0                        | 766 (76.6%)         | IIIIIIIIIIIIIII
## |     | [factor]         | 2. 1                        | 234 (23.4%)         | IIII
## +-----+------------------+-----------------------------+---------------------+-----------------------
## | 6   | USED_CAR         | 1. 0                        | 897 (89.7%)         | IIIIIIIIIIIIIIIII
## |     | [factor]         | 2. 1                        | 103 (10.3%)         | II
## +-----+------------------+-----------------------------+---------------------+-----------------------
## | 7  | FURNITURE        | 1. 0                        | 819 (81.9%)         | IIIIIIIIIIIIIIII
## |    | [factor]         | 2. 1                        | 181 (18.1%)         | III
## +----+------------------+-----------------------------+---------------------+----------------------
## | 8  | RADIO.TV         | 1. 0                        | 720 (72.0%)         | IIIIIIIIIIIIII
## |    | [factor]         | 2. 1                        | 280 (28.0%)         | IIIII
## +----+------------------+-----------------------------+---------------------+----------------------
## | 9  | EDUCATION        | 1. 0                        | 950 (95.0%)         | IIIIIIIIIIIIIIIIIII
## |    | [factor]         | 2. 1                        |  50 ( 5.0%)         | I
## +----+------------------+-----------------------------+---------------------+----------------------
## | 10 | RETRAINING       | 1. 0                        | 903 (90.3%)         | IIIIIIIIIIIIIIIIII
## |    | [factor]         | 2. 1                        |  97 ( 9.7%)         | I
## +----+------------------+-----------------------------+---------------------+----------------------
## | 11 | AMOUNT           | Mean (sd) : 3271.3 (2822.7) | 921 distinct values | :
## |    | [numeric]        | min < med < max:            |                     | : .
## |    |                  | 250 < 2319.5 < 18424        |                     | : :
## |    |                  | IQR (CV) : 2606.8 (0.9)     |                     | : :
## |    |                  |                             |                     | : : : : .
## +----+------------------+-----------------------------+---------------------+----------------------
## | 12 | SAV_ACCT         | 1. 0                        | 603 (60.3%)         | IIIIIIIIIIII
## |    | [factor]         | 2. 1                        | 103 (10.3%)         | II
## |    |                  | 3. 2                        |  63 ( 6.3%)         | I
## |    |                  | 4. 3                        |  48 ( 4.8%)         |
## |    |                  | 5. 4                        | 183 (18.3%)         | III
## +----+------------------+-----------------------------+---------------------+----------------------
## | 13 | EMPLOYMENT       | 1. 0                        |  62 ( 6.2%)         | I
## |    | [factor]         | 2. 1                        | 172 (17.2%)         | III
## |    |                  | 3. 2                        | 339 (33.9%)         | IIIIII
## |    |                  | 4. 3                        | 174 (17.4%)         | III
## |    |                  | 5. 4                        | 253 (25.3%)         | IIIII
## +----+------------------+-----------------------------+---------------------+----------------------
## | 14 | INSTALL_RATE     | Mean (sd) : 3 (1.1)         | 1 : 136 (13.6%)     | II
## |    | [numeric]        | min < med < max:            | 2 : 231 (23.1%)     | IIII
## |    |                  | 1 < 3 < 4                   | 3 : 157 (15.7%)     | III
## |    |                  | IQR (CV) : 2 (0.4)          | 4 : 476 (47.6%)     | IIIIIIIII
## +----+------------------+-----------------------------+---------------------+----------------------
## | 15 | MALE_DIV         | 1. 0                        | 950 (95.0%)         | IIIIIIIIIIIIIIIIIII
## |    | [factor]         | 2. 1                        |  50 ( 5.0%)         | I
## +----+------------------+-----------------------------+---------------------+----------------------
## | 16 | MALE_SINGLE      | 1. 0                        | 452 (45.2%)         | IIIIIIIII
## |    | [factor]         | 2. 1                        | 548 (54.8%)         | IIIIIIIIII
## +----+------------------+-----------------------------+---------------------+----------------------
## | 17 | MALE_MAR_or_WID  | 1. 0                        | 908 (90.8%)         | IIIIIIIIIIIIIIIIII
## |    | [factor]         | 2. 1                        |  92 ( 9.2%)         | I
## +----+------------------+-----------------------------+---------------------+----------------------
## | 18 | CO.APPLICANT     | 1. 0                        | 959 (95.9%)         | IIIIIIIIIIIIIIIIIII
## |    | [factor]         | 2. 1                        |  41 ( 4.1%)         |
## +----+------------------+-----------------------------+---------------------+----------------------
## | 19 | GUARANTOR        | 1. 0                        | 948 (94.8%)         | IIIIIIIIIIIIIIIIII
## |    | [factor]         | 2. 1                        |  52 ( 5.2%)         | I
## +----+------------------+-----------------------------+---------------------+----------------------
## | 20 | PRESENT_RESIDENT | 1. 1                        | 130 (13.0%)         | II
## |    | [factor]         | 2. 2                        | 308 (30.8%)         | IIIIII
## |    |                  | 3. 3                        | 149 (14.9%)         | II
## |    |                  | 4. 4                        | 413 (41.3%)         | IIIIIIII
## +----+------------------+-----------------------------+---------------------+------------------------
## | 21 | REAL_ESTATE      | 1. 0                        | 718 (71.8%)         | IIIIIIIIIIIIII
## |    | [factor]         | 2. 1                        | 282 (28.2%)         | IIIII
## +----+------------------+-----------------------------+---------------------+------------------------
## | 22 | PROP_UNKN_NONE   | 1. 0                        | 846 (84.6%)         | IIIIIIIIIIIIIIII
## |    | [factor]         | 2. 1                        | 154 (15.4%)         | III
## +----+------------------+-----------------------------+---------------------+------------------------
## | 23 | AGE              | Mean (sd) : 35.5 (11.4)     | 53 distinct values  |   :
## |    | [numeric]        | min < med < max:            |                     |   : .
## |    |                  | 19 < 33 < 75                |                     | : : : :
## |    |                  | IQR (CV) : 15 (0.3)         |                     | : : : : :
## |    |                  |                             |                     | : : : : : : : . .
## +----+------------------+-----------------------------+---------------------+------------------------
## | 24 | OTHER_INSTALL    | 1. 0                        | 814 (81.4%)         | IIIIIIIIIIIIIIII
## |    | [factor]         | 2. 1                        | 186 (18.6%)         | III
## +----+------------------+-----------------------------+---------------------+------------------------
## | 25 | RENT             | 1. 0                        | 821 (82.1%)         | IIIIIIIIIIIIIIII
## |    | [factor]         | 2. 1                        | 179 (17.9%)         | III
## +----+------------------+-----------------------------+---------------------+------------------------
## | 26 | OWN_RES          | 1. 0                        | 287 (28.7%)         | IIIII
## |    | [factor]         | 2. 1                        | 713 (71.3%)         | IIIIIIIIIIIIII
## +----+------------------+-----------------------------+---------------------+------------------------
## | 27 | NUM_CREDITS      | Mean (sd) : 1.4 (0.6)       | 1 : 633 (63.3%)     | IIIIIIIIIIII
## |    | [numeric]        | min < med < max:            | 2 : 333 (33.3%)     | IIIIII
## |    |                  | 1 < 1 < 4                   | 3 :  28 ( 2.8%)     |
## |    |                  | IQR (CV) : 1 (0.4)          | 4 :   6 ( 0.6%)     |
## +----+------------------+-----------------------------+---------------------+------------------------
## | 28 | JOB              | 1. 0                        |  22 ( 2.2%)         |
## |    | [factor]         | 2. 1                        | 200 (20.0%)         | IIII
## |    |                  | 3. 2                        | 630 (63.0%)         | IIIIIIIIIIII
## |    |                  | 4. 3                        | 148 (14.8%)         | II
## +----+------------------+-----------------------------+---------------------+------------------------
## | 29 | NUM_DEPENDENTS   | Min  : 1                    | 1 : 845 (84.5%)     | IIIIIIIIIIIIIIII
## |    | [numeric]        | Mean : 1.2                  | 2 : 155 (15.5%)     | III
## |    |                  | Max  : 2                    |                     |
## +----+------------------+-----------------------------+---------------------+------------------------
## | 30 | TELEPHONE        | 1. 0                        | 596 (59.6%)         | IIIIIIIIIII
## |    | [factor]         | 2. 1                        | 404 (40.4%)         | IIIIIIII
## +----+------------------+-----------------------------+---------------------+------------------------
## | 31 | FOREIGN          | 1. 0                        | 963 (96.3%)         | IIIIIIIIIIIIIIIIIII
## |    | [factor]         | 2. 1                        |  37 ( 3.7%)         |
## +----+------------------+-----------------------------+---------------------+------------------------
## | 32 | RESPONSE         | 1. 0                        | 300 (30.0%)         | IIIIII
## |    | [factor]         | 2. 1                        | 700 (70.0%)         | IIIIIIIIIIIIII
## +----+------------------+-----------------------------+---------------------+------------------------
```

Correlation plot :

Correlation plot between continuous variables :

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

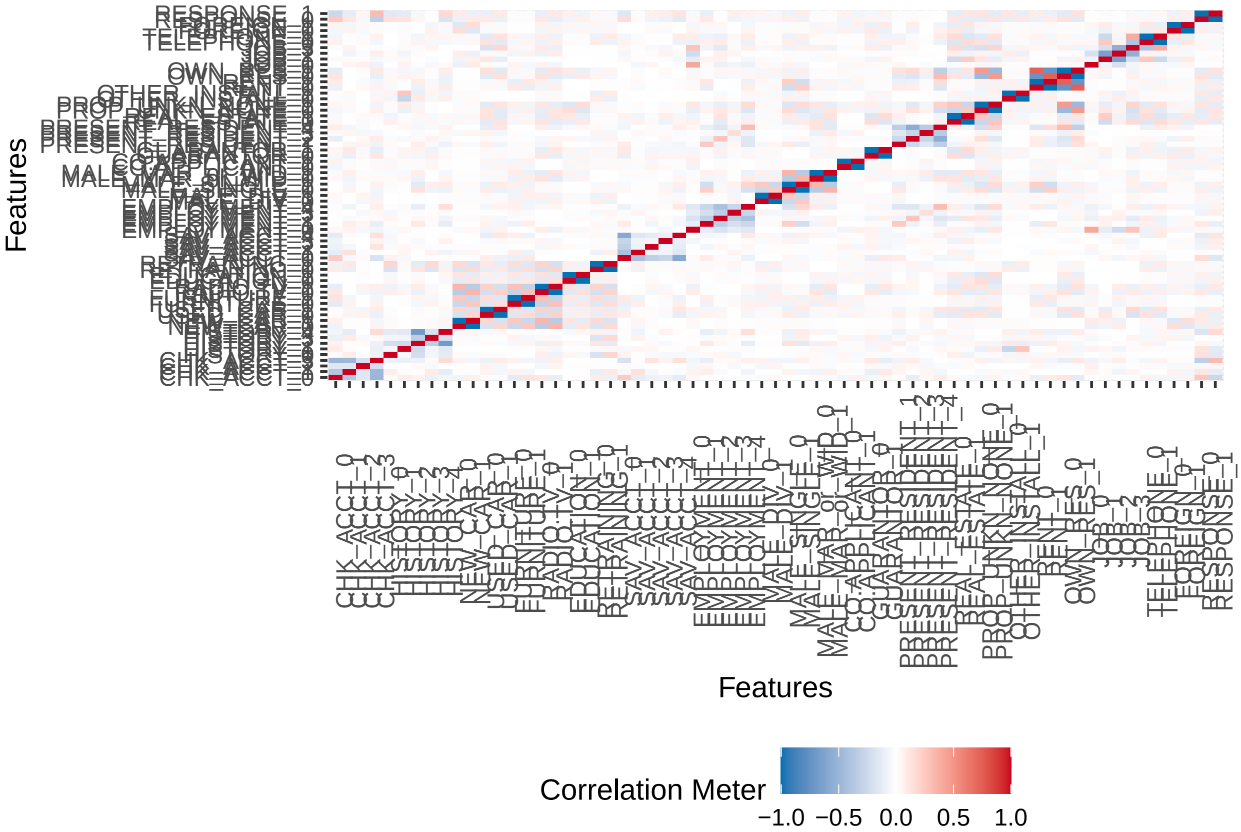


There are little correlation between the continuous variables. We can notice that there is a correlation of 62% between the variable **DURATION** and **AMOOUNT**. This makes sense and is known by the bankers that the bigger the amount of credit, the longer the duration of the credit in months will last.

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)

PCA Exploration:

It is good to perform a PCA Exploration in order to reduce the dimensions or/and see which variables to select.

We start by selecting the numerical values:

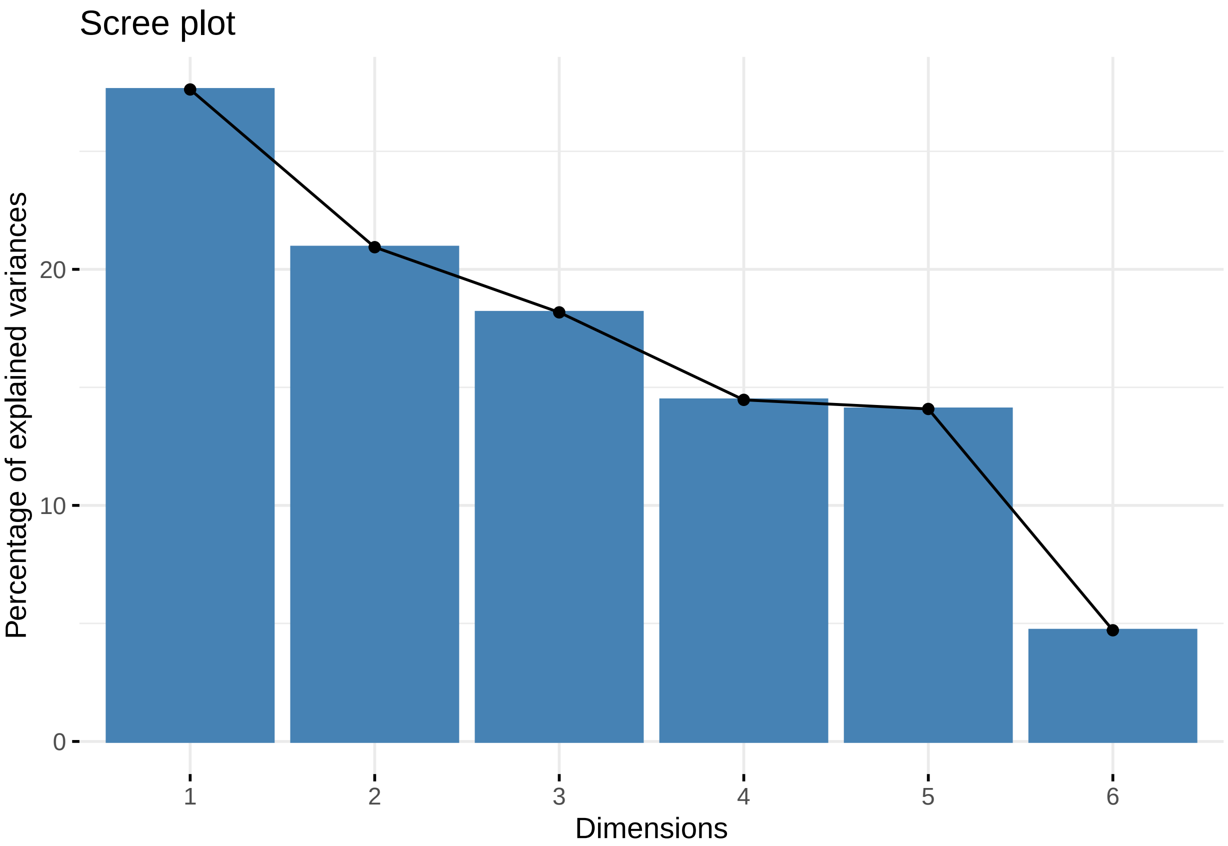
```r
German_credit.num <- German_credit %>%
  select('DURATION', 'AMOUNT', 'INSTALL_RATE', 'AGE', 'NUM_CREDITS','NUM_DEPENDENTS')
```

```r
German_credit.pca <- prcomp(German_credit.num, center = TRUE, scale = TRUE)
summary(German_credit.pca)
```

```
## Importance of components:
##                           PC1    PC2    PC3    PC4    PC5     PC6
## Standard deviation     1.2873 1.1208 1.0443 0.9318 0.9193 0.53164
## Proportion of Variance 0.2762 0.2094 0.1818 0.1447 0.1409 0.04711
## Cumulative Proportion  0.2762 0.4856 0.6673 0.8120 0.9529 1.00000
```

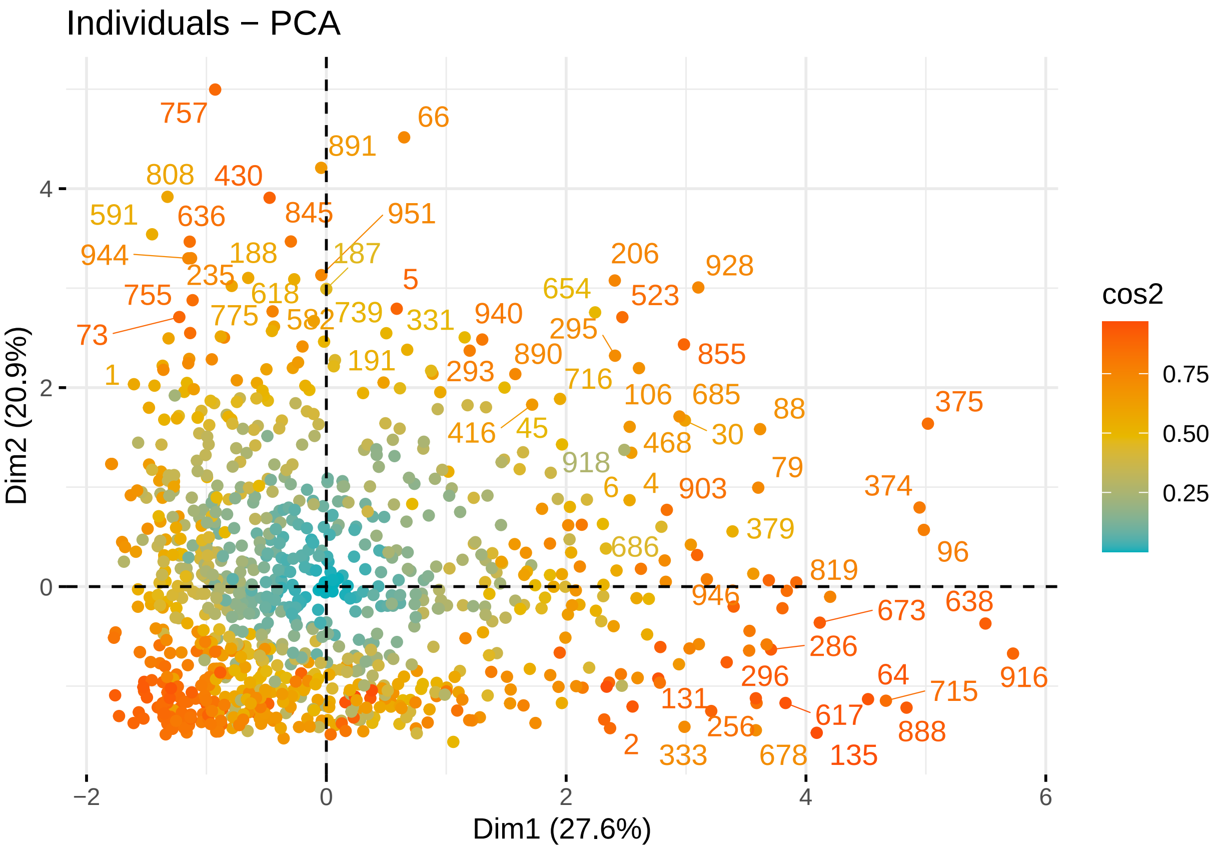
From the PCA summary we can see 4 principal components should be taken into account in order to explain approximately 81% of the variation of the data.

```
fviz_eig(German_credit.pca)
```

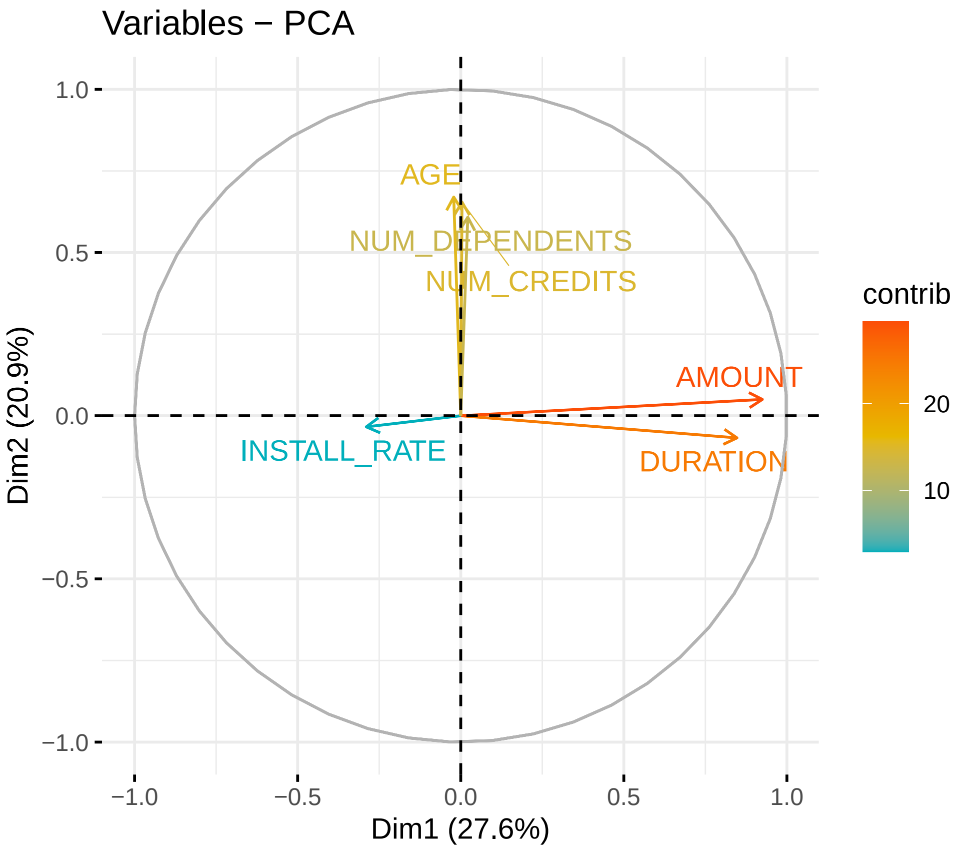


```
fviz_pca_ind(German_credit.pca,
             col.ind = "cos2", # Colorer par le cos2
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE
             )
```

```
## Warning: ggrepel: 933 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



```
fviz_pca_var(German_credit.pca,
             col.var = "contrib",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE
             )
```



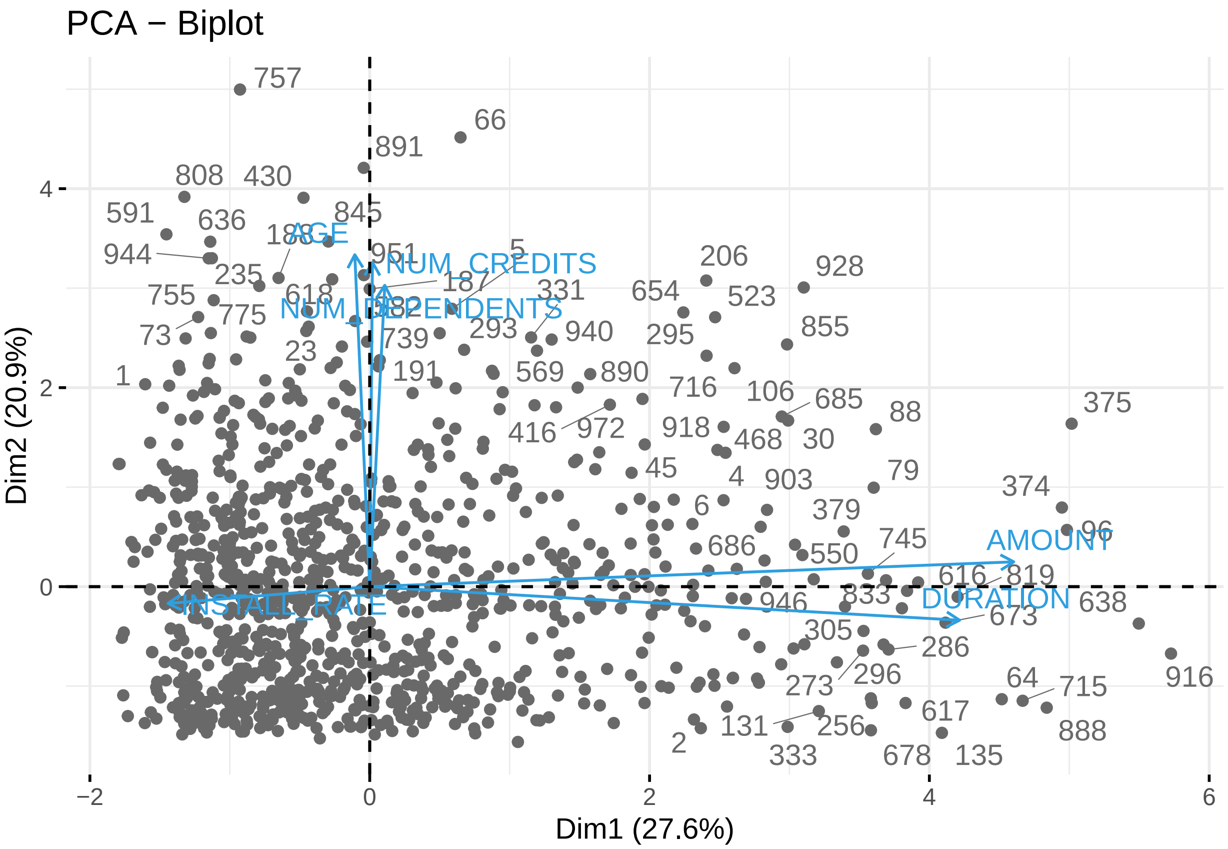
From this circle of correlations, we see that :

- The first principal component PC1 is strongly positively correlated with the variables **AMOUNT** and **DURATION**. So the larger PC1, the larger these features. It is also a little bit negatively correlated with **INSTALL_RATE**.

- The second principal component PC2 is strongly positively correlated with **AGE**, **NUM_DEPENDENTS** and **NUM_CREDITS**.

```
fviz_pca_biplot(German_credit.pca, repel = TRUE,
                col.var = "#2E9FDF",
                col.ind = "#696969"
                )
```

```
## Warning: ggrepel: 924 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



From this biplot, we can see some characteristics of the observations.

We can export the dataset as we have made some modifications. It will be easier for the other files.

```
# write.csv(German_credit,"./../Data_DA/GermanCredit_cleaned.csv", row.names = FALSE)
```