German Credit Dataset

1 Introduction

German Credit Dataset is a research dataset from the University of Hamburg from 1994 and donated by Prof. Hans Hoffman. Each entry represents a person who takes a credit by a bank. Each person is classified as “good” or “bad” credit risks according to the set of attributes.

Figure 1: Source: freepik (link)

Dataset basic information:

Variable	Description
`class` (target)	“good”/“bad”
checking_status	Status of the existing checking account, in Deutsche Mark
duration	Duration in months
credit_history	Credit history
credit_amount	Amount of the desired credit
saving_status	Status of savings account/bonds, in Deutsche Mark
employment	Present employment, in number of years
installment_commitment	Installment rate in percentage of disposable income
personal_status	Personal status and sex

Variable	Description
other_parties	Other debtors or guarantors
residence_since	Current residence since, in years
age	Age in years
other_payment_plans	Other installment plans
existing_credits	Number of existing credits at this bank
job	Current job
num_dependents	Number of people being liable to provide maintenance for
own_telephone	Telephone ("yes"/"none")
foreign_worker	Foreign worker ("yes"/"no")

We use OpenML (R-Package) to download the dataset in a machine-readable format and convert it into a `data.frame`:

```r
# load the dataset from OpenML Library
d <- OpenML::getOMLDataSet(data.id = 31)

# convert the OpenML object to a tibble (enhanced data.frame)
credit <- d %>% dplyr::as_tibble()
skimmed_credit <- skimr::skim(credit)
print(credit, width = Inf)
```

```
## # A tibble: 1,000 x 21
##    checking_status duration credit_history                 purpose
##    <fct>              <dbl> <fct>                          <fct>
##  1 <0                     6 critical/other existing credit radio/tv
##  2 0<=X<200              48 existing paid                  radio/tv
##  3 no checking           12 critical/other existing credit education
##  4 <0                    42 existing paid                  furniture/equipment
##  5 <0                    24 delayed previously             new car
##  6 no checking           36 existing paid                  education
##  7 no checking           24 existing paid                  furniture/equipment
##  8 0<=X<200              36 existing paid                  used car
##  9 no checking           12 existing paid                  radio/tv
## 10 0<=X<200              30 critical/other existing credit new car
##    credit_amount savings_status   employment installment_commitment
##            <dbl> <fct>            <fct>                       <dbl>
##  1          1169 no known savings >=7                             4
##  2          5951 <100             1<=X<4                          2
##  3          2096 <100             4<=X<7                          2
##  4          7882 <100             4<=X<7                          2
##  5          4870 <100             1<=X<4                          3
##  6          9055 no known savings 1<=X<4                          2
##  7          2835 500<=X<1000      >=7                             3
##  8          6948 <100             1<=X<4                          2
##  9          3059 >=1000           4<=X<7                          2
## 10          5234 <100             unemployed                      4
##    personal_status    other_parties residence_since property_magnitude   age
##    <fct>              <fct>                   <dbl> <fct>              <dbl>
##  1 male single        none                        4 real estate           67
##  2 female div/dep/mar none                        2 real estate           22
##  3 male single        none                        3 real estate           49
```

```
##  4 male single       guarantor                   4 life insurance       45
##  5 male single       none                        4 no known property    53
##  6 male single       none                        4 no known property    35
##  7 male single       none                        4 life insurance       53
##  8 male single       none                        2 car                  35
##  9 male div/sep      none                        4 real estate          61
## 10 male mar/wid      none                        2 car                  28
##    other_payment_plans housing  existing_credits job
##    <fct>               <fct>               <dbl> <fct>
##  1 none                own                     2 skilled
##  2 none                own                     1 skilled
##  3 none                own                     1 unskilled resident
##  4 none                for free                1 skilled
##  5 none                for free                2 skilled
##  6 none                for free                1 unskilled resident
##  7 none                own                     1 skilled
##  8 none                rent                    1 high qualif/self emp/mgmt
##  9 none                own                     1 unskilled resident
## 10 none                own                     2 high qualif/self emp/mgmt
##    num_dependents own_telephone foreign_worker class
##             <dbl> <fct>         <fct>          <fct>
##  1              1 yes           yes            good
##  2              1 none          yes            bad
##  3              2 none          yes            good
##  4              2 none          yes            good
##  5              2 none          yes            bad
##  6              2 yes           yes            good
##  7              1 none          yes            good
##  8              1 yes           yes            good
##  9              1 none          yes            good
## 10              1 none          yes            bad
## # ... with 990 more rows
```

2 Exploratory Data Analysis (EDA)

In this part, we will walk through a few characteristics of credit dataset using library `skimr` and `DataExplorer`.

2.1 Factor variables

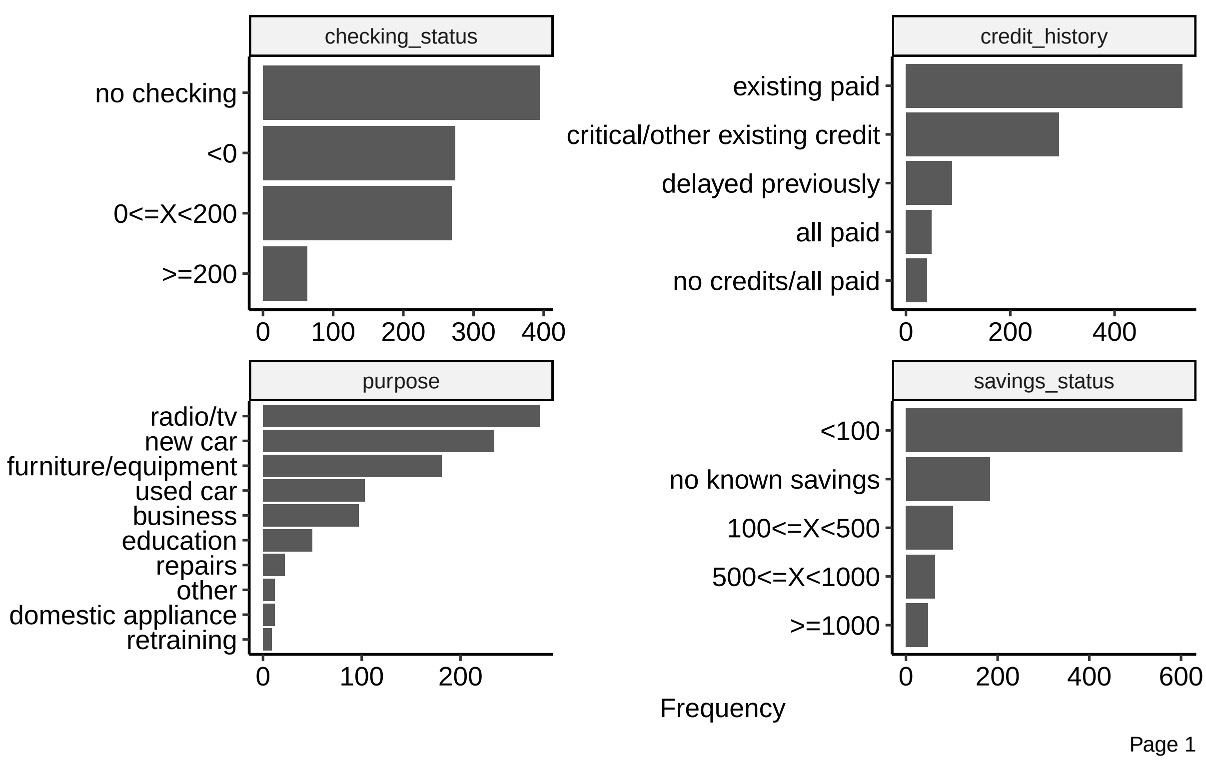
General statistics about factor variables from credit dataset:

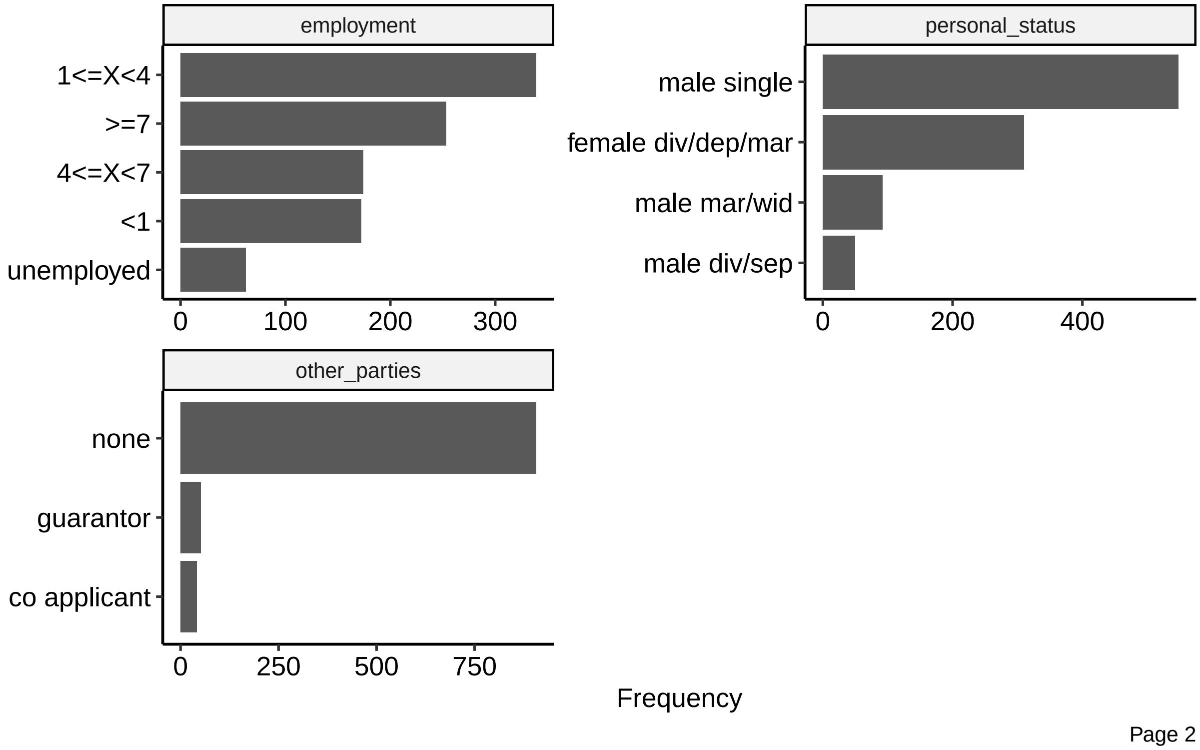
```
skimr::partition(skimmed_credit)$factor %>%
        knitr::kable(format = 'latex', booktabs = TRUE) %>%
        kableExtra::kable_styling(latex_options = 'HOLD_position')
```

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
checking_status	0	1	FALSE	4	no : 394, <0: 274, 0<=: 269, >=2: 63
credit_history	0	1	FALSE	5	exi: 530, cri: 293, del: 88, all: 49
purpose	0	1	FALSE	10	rad: 280, new: 234, fur: 181, use: 103
savings_status	0	1	FALSE	5	<10: 603, no : 183, 100: 103, 500: 63
employment	0	1	FALSE	5	1<=: 339, >=7: 253, 4<=: 174, <1: 172
personal_status	0	1	FALSE	4	mal: 548, fem: 310, mal: 92, mal: 50
other_parties	0	1	FALSE	3	non: 907, gua: 52, co : 41
property_magnitude	0	1	FALSE	4	car: 332, rea: 282, lif: 232, no : 154
other_payment_plans	0	1	FALSE	3	non: 814, ban: 139, sto: 47
housing	0	1	FALSE	3	own: 713, ren: 179, for: 108
job	0	1	FALSE	4	ski: 630, uns: 200, hig: 148, une: 22
own_telephone	0	1	FALSE	2	non: 596, yes: 404
foreign_worker	0	1	FALSE	2	yes: 963, no: 37
class	0	1	FALSE	2	goo: 700, bad: 300

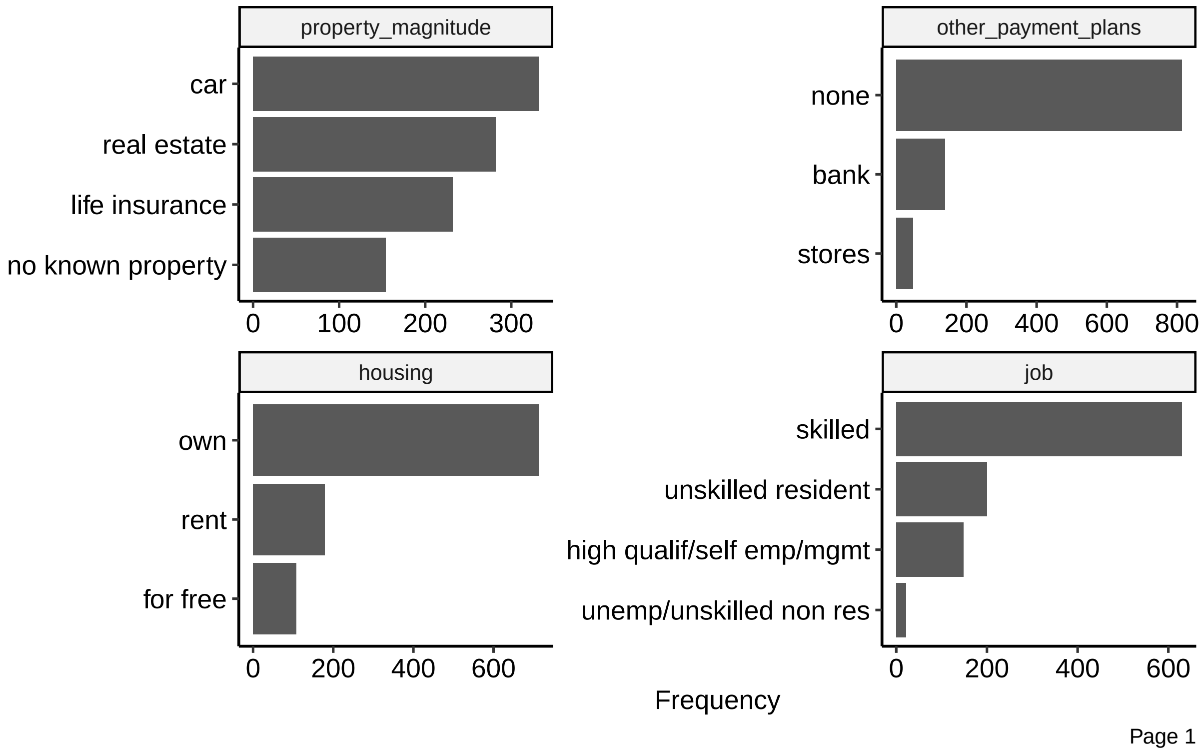
From the general statistics, it can be seen that there is no missing value. The majority of factor variables has fewer than 5 unique values, the exceptions are `credit_history`, `saving_status`, `employment` with 5 unique values and `purpose` with 10 unique values.

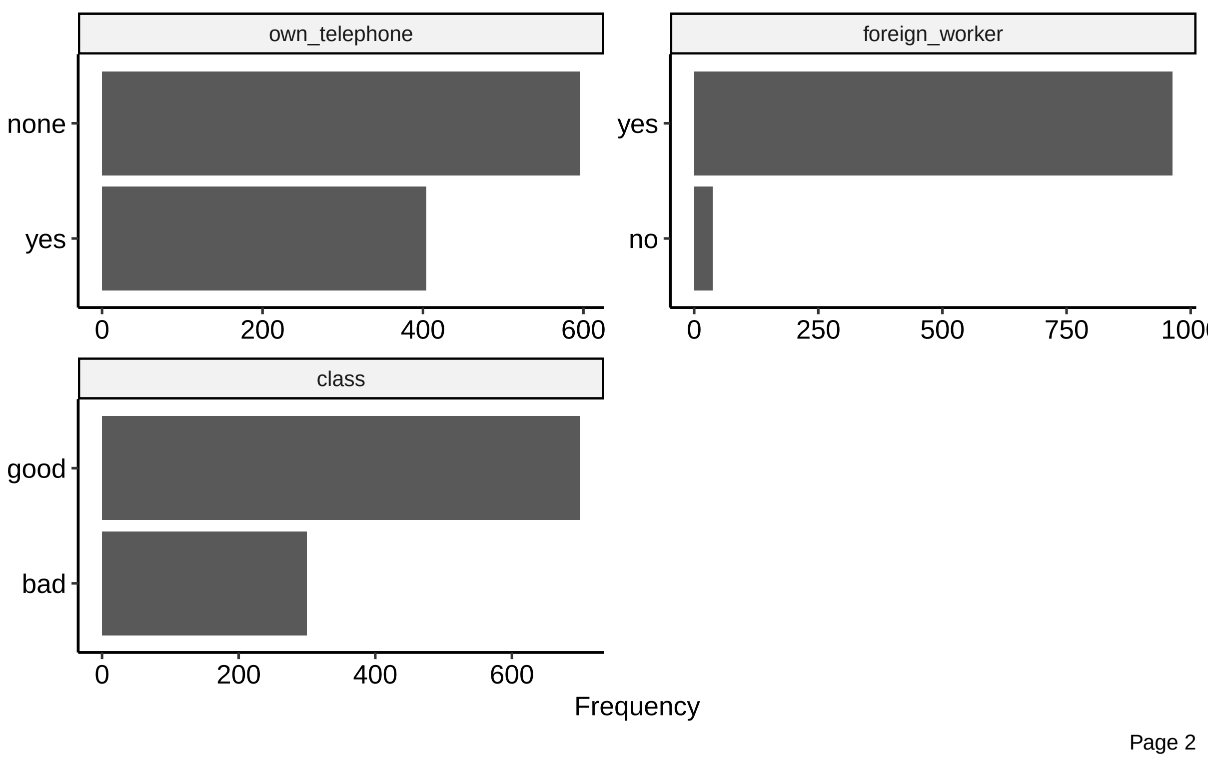
```
credit_factor <- credit %>% select(where(is.factor))
DataExplorer::plot_bar(
        credit_factor %>% select(1:7),
        ggtheme = ggpubr::theme_pubr(base_size = 10),
        nrow = 2,
        ncol = 2
)
```





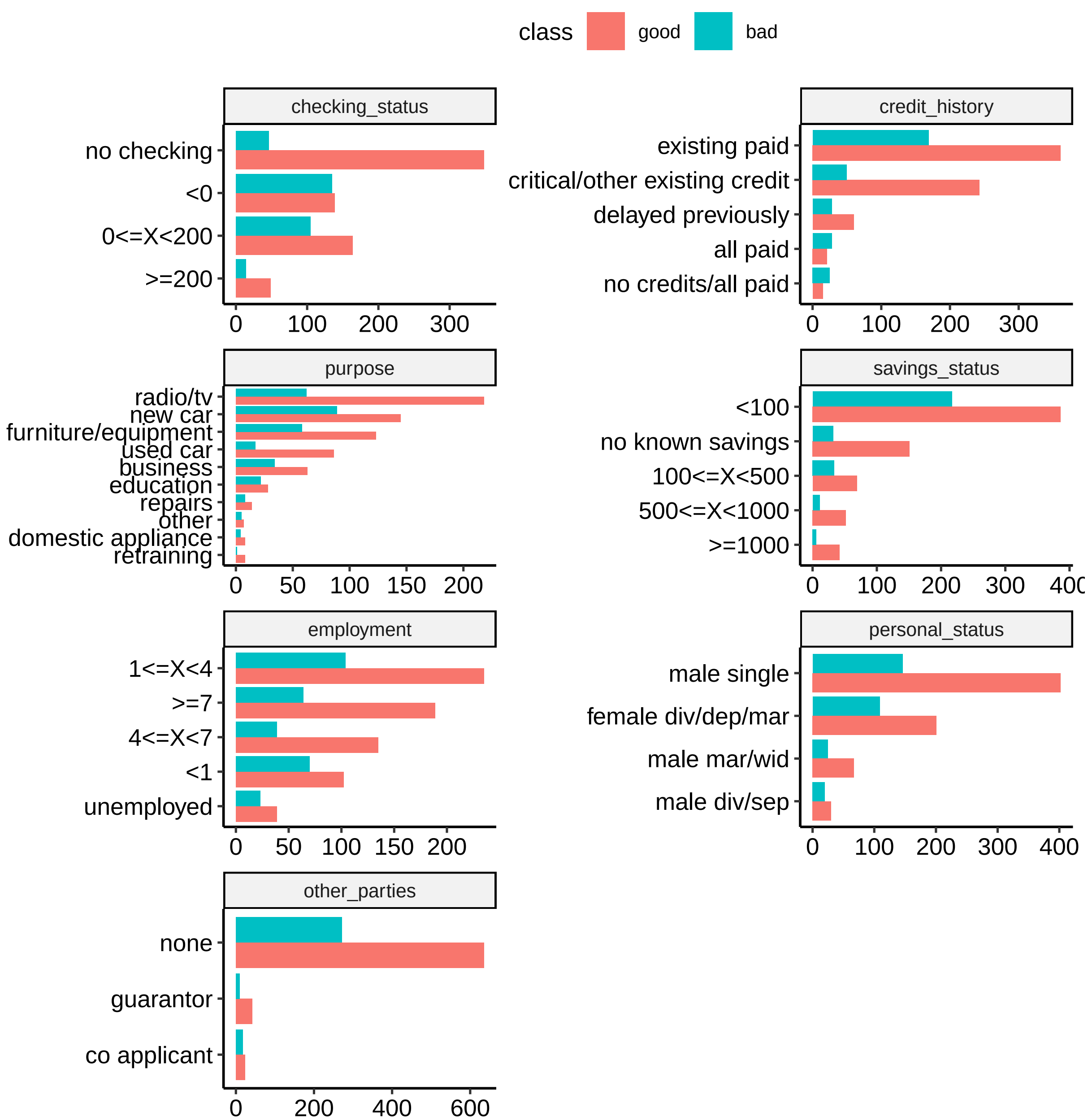

```
DataExplorer::plot_bar(
        credit_factor %>% select(8:14),
        ggtheme = ggpubr::theme_pubr(base_size = 10),
        nrow = 2,
        ncol = 2
)
```



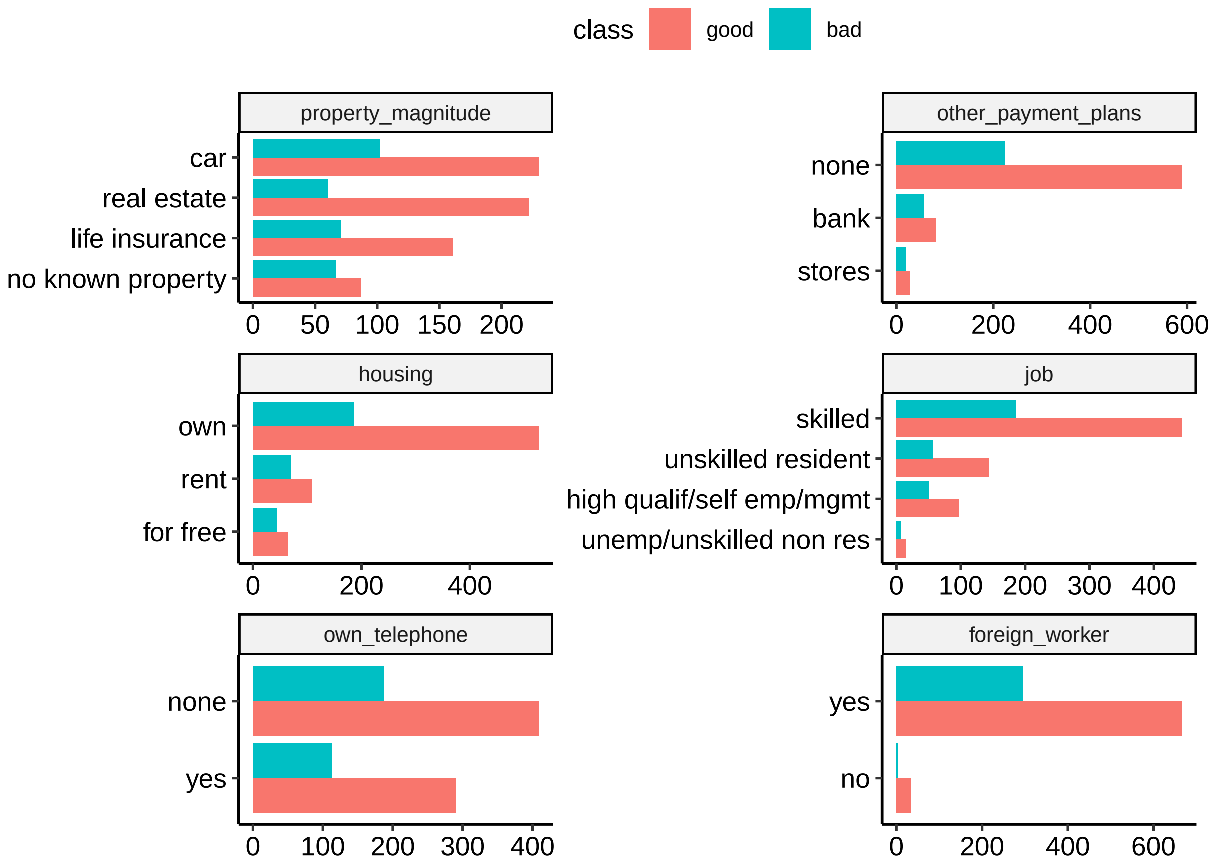


According to the bar plots, it can be seen that most of the factor variables are imbalanced, including the response variable `class`. The majority of data points in this dataset has good credit (70%). Foreign workers account for 96.3% of the whole dataset. 71.3% owns real estate and 63% of the data comes from skilled workers. The majority of people in the dataset applies for the credit to buy appliances and new cars: radio/tv (28%), new car (23.4%) and furniture (18.1%). Notably, more than 90% of the records apply for credit alone without other guarantor or co applicant.

```
DataExplorer::plot_bar(
        credit_factor %>% select(c(1:7, class)),
        by = "class",
        by_position = "dodge",
        ggtheme = ggpubr::theme_pubr(base_size = 10),
        nrow = 4,
        ncol = 2
)
```



```
DataExplorer::plot_bar(
        credit_factor %>% select(8:14),
        by = "class",
        by_position = "dodge",
        ggtheme = ggpubr::theme_pubr(base_size = 10),
        ncol = 2,
        nrow = 4
)
```



After taking the class into consideration, the one thing that stands out is that for every unique value of each factor variable, the proportion of people who have good credit is always higher than the ones having bad credit.

2.2 Numerical variables

General statistics about numerical variables from credit dataset:

```
skimmed_credit_numeric <- skimr::partition(skimmed_credit)$numeric
split(1:ncol(skimmed_credit_numeric),
      sort(rep_len(1:2, ncol(skimmed_credit_numeric)))) %>%
        map(~select(skimmed_credit_numeric, .)) %>%
        map(knitr::kable, booktabs = T) %>%
```

```
  map(kableExtra::kable_styling, latex_options = 'HOLD_position') %>%
  walk(print)
```

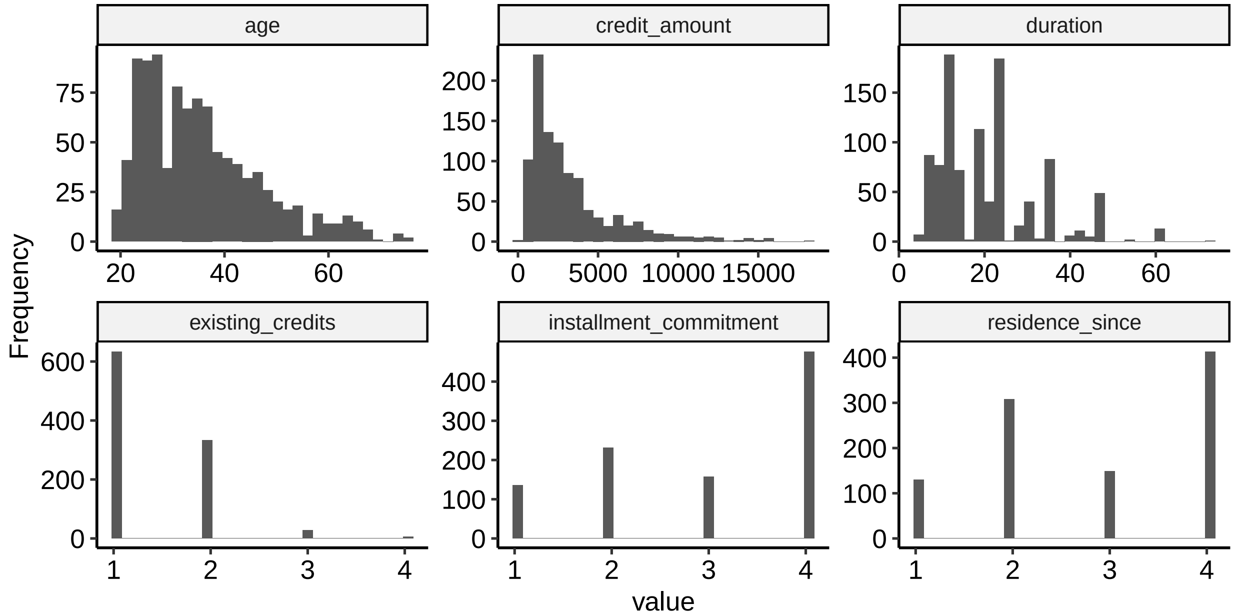
skim_variable	n_missing	complete_rate	mean	sd	p0
duration	0	1	20.903	12.0588145	4
credit_amount	0	1	3271.258	2822.7368760	250
installment_commitment	0	1	2.973	1.1187147	1
residence_since	0	1	2.845	1.1037179	1
age	0	1	35.546	11.3754686	19
existing_credits	0	1	1.407	0.5776545	1
num_dependents	0	1	1.155	0.3620858	1

p25	p50	p75	p100	hist
12.0	18.0	24.00	72	
1365.5	2319.5	3972.25	18424	
2.0	3.0	4.00	4	
2.0	3.0	4.00	4	
27.0	33.0	42.00	75	
1.0	1.0	2.00	4	
1.0	1.0	1.00	2	

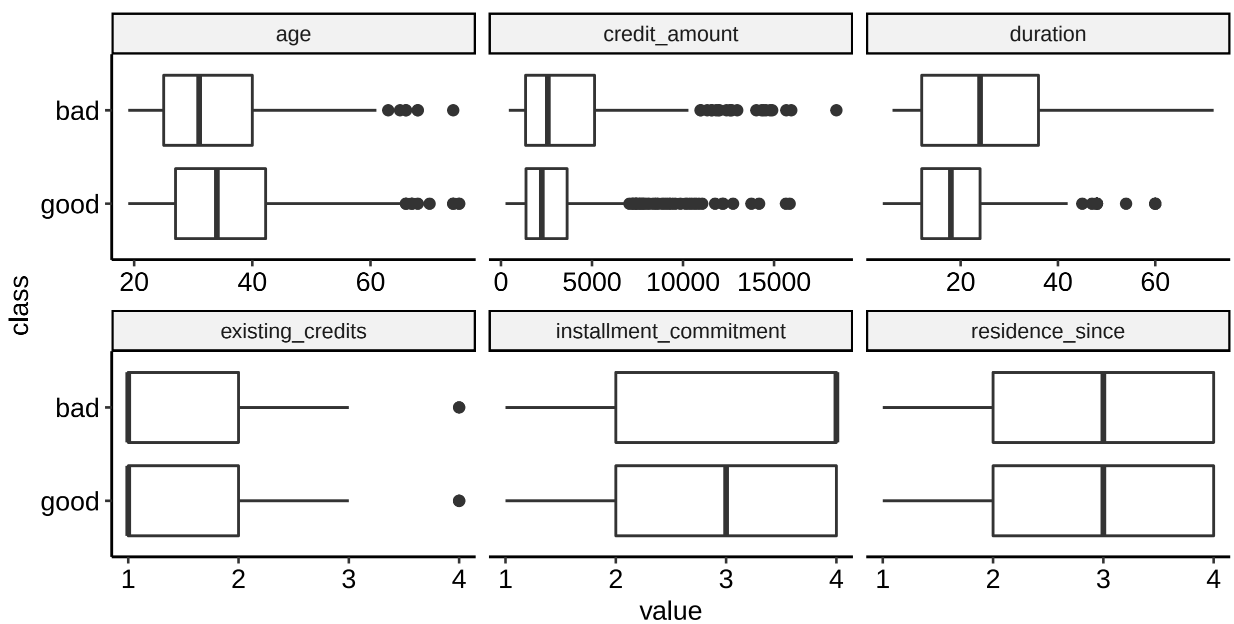
As can be seen from the statistics, similar to the factor variables, numerical variables in this dataset don't have missing values. The ranges of values of the numerical features extremely differ from one to another.

To have a better view at the distributions of these features, let's take a look at their histograms and their boxplots (broken down by class labels).

```
DataExplorer::plot_histogram(
        credit,
        ggtheme = ggpubr::theme_pubr(base_size = 10),
        ncol = 3, nrow = 2)
```




```
DataExplorer::plot_boxplot(
        credit,
        by = "class",
        ggtheme = ggpubr::theme_pubr(base_size = 10),
        ncol = 3, nrow = 2)
```

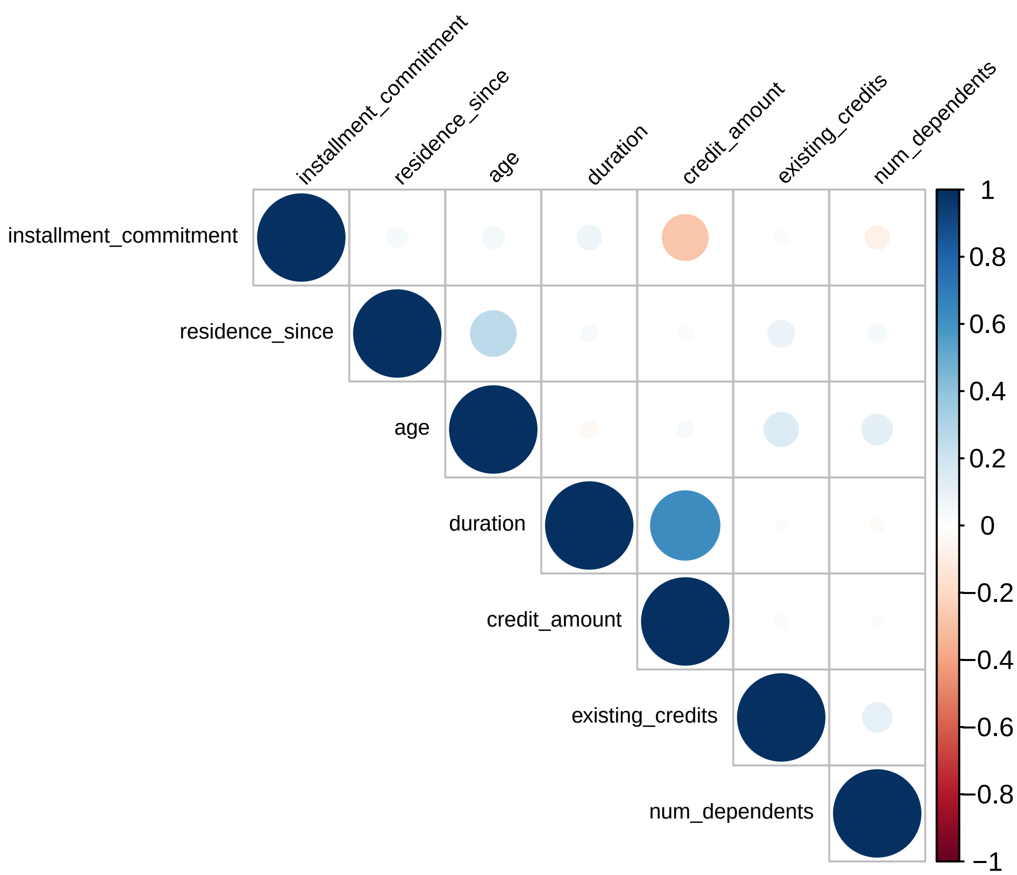


Three numerical variables `age`, `credit_amount` and `duration` have right-skewed distribution. The majority of the age of people applying for credit varies within the range from 25-40. The usual credit amount in this dataset is less than 4000. The typical duration that people from this dataset applied for is less than 2 years.

From the boxplots, it can be seen that there is no visible strong connection between the numerical variables and the response.

To understand more the linear relationship between the pairs of numerical variables, we create a correlation matrix:

```
credit %>% select(where(is.numeric)) %>%
        cor() %>%
        corrplot(
           type = "upper",
           order = "hclust",
           tl.col = "black",
           tl.srt = 45,
           tl.cex = 0.7
        )
```



According to the correlation plot, there is a notable positive correlation between `duration` and `credit_amount`. This makes much sense in real life as the more money people loan, the more time they need to pay it back.