Predicting Individual Good or Bad Credit Risk

Machine Learning - Applied Project - Phase I

Saurabh Mallik (s3623575) & Dilip Chandra (s3574580)

Introduction

“It takes money to make more money”. As an individual or company when we want to lend money, we set some critical parameters or guidelines to understand the credit risk. In this project, our aim is to analyze good and bad credit risk associated with individuals. The purpose of this stage of the project is to build classifiers which will help in predicting whether or not an individual has good or bad credit risk. This will be based on German Credit Dataset, which was sourced from UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)). The project is segregated into two distinct stages. Stage 1 focusses on creating classifiers which will help the machine learning process to provide prediction. This stage will be comprise of data preprocessing, exploration and visualisation. In stage 2, we will focus on model building. The report is divided into the following parts:

1. Dataset: Which talks about the data
2. Pre-Processing: Which takes one through the required steps to clean and tidy the data for exploration.
3. Exploration: Where we show the dependencies and key features of the data
4. Summary: Where we summarise the report and findings.

Dataset

The German Credit Dataset sourced from UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)) provides us with two datasets. For this research we will be using the original dataset in categorical form provided by Prof. Hofman. The dataset has 1000 observations and 21 variables. The dataset consists of 20 descriptive feature and 1 target feature. The dataset will be used in stage 2 to build classifiers and evaluate the execution and fitting of the models using cross-validation.

Target Feature

The response feature for this project is CreditRisk which has two classes good or bad, and is hence a binary classification problem. The goal of this project is to predict whether an individual had good credit or bad credit risk.

Descriptive Features

The variable descriptions for the dataset are given below:

1. Checking: <0, 0<X<200, >=200, no checking.
2. Duration: continuous.
3. CreditHistory: all paid, eisting paid, delayed previously, critical/other existing credit, no credits/all paid.
4. Purpose: business, domestic appliance, education furniture/equipment, new car, other, radio/tv, repairs, retraining, used car.
5. CreditAmount: continuous.
6. Saving: <100, >=1000, 100<=X<500, 500<=X<1000, no known savings.
7. Employment: <1, >=7, 1<=X<4, 4<=X<7, unemployed.

8. Installment: continuous.
9. Status: male div/sep, male mar/wid, male single, female div/dep/mar.
10. OtherParties: co applicant, guarantor, none.
11. Residence: continuous.
12. Asset: life insurance, no known property, real estate, car.
13. Age: continuous.
14. OtherPlans: bank, none, stores.
15. Housing: for free, own, rent.
16. ExistingCredits: continuous.
17. JobType: high qualif/self emp/mgmt, unemp/unskilled non res, unskilled resident, skilled.
18. Dependants: continuous.
19. Phone: none, yes.
20. Foreign: no, yes.

The descriptors are quite self explanatory and hence comprehension for exploration was suitable.

Data Pre-processing

Preliminaries

The following R packages were used for this stage of the project.

```
library(readr)
library(ggplot2)
library(dplyr)
library(mlr)
library(tidyverse)
library(GGally)
library(cowplot)
```

We read the dataset and renamed the variables for convenience and ease of usage as there were 20 distinct variables in the dataset. We eventually convert the character variables to factors (categorical data).

```
credit <- read_csv("credit.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   duration = col_integer(),
##   credit_amount = col_integer(),
##   installment_commitment = col_integer(),
##   residence_since = col_integer(),
##   age = col_integer(),
##   existing_credits = col_integer(),
##   num_dependents = col_integer()
## )
```

```
## See spec(...) for full column specifications.
```

```
names(credit) <- c('Checking', 'Duration', 'CreditHistory', 'Purpose', 'CreditAmount'
,
                   'Saving', 'Employment', 'Installment', 'Status', 'OtherParties',
'Residence',
                   'Asset', 'Age', 'OtherPlans', 'Housing', 'ExistingCredits', 'JobT
ype', 'Dependants',
                  'Phone', 'Foreign', 'CreditRisk')
```

Tidying the dataset

We use the str and summarizeColumns functions to understand the dataset a bit further. From these we noticed the following:
1. All the character variables had a lot of white space.
2. There were no missing values in the dataset.

```
str(credit, give.attr = FALSE)
```

```
## tibble [1,000 × 21] (S3: tbl_df/tbl/data.frame)
##  $ Checking       : chr [1:1000] "'<0'" "'0<=X<200'" "'no checking'" "'<0'" ...
##  $ Duration       : int [1:1000] 6 48 12 42 24 36 24 36 12 30 ...
##  $ CreditHistory  : chr [1:1000] "'critical/other existing credit'" "'existing pai
d'" "'critical/other existing credit'" "'existing paid'" ...
##  $ Purpose        : chr [1:1000] "radio/tv" "radio/tv" "education" "furniture/equi
pment" ...
##  $ CreditAmount   : int [1:1000] 1169 5951 2096 7882 4870 9055 2835 6948 3059 5234
...
##  $ Saving         : chr [1:1000] "'no known savings'" "'<100'" "'<100'" "'<100'"
...
##  $ Employment     : chr [1:1000] "'>=7'" "'1<=X<4'" "'4<=X<7'" "'4<=X<7'" ...
##  $ Installment    : int [1:1000] 4 2 2 2 3 2 3 2 2 4 ...
##  $ Status         : chr [1:1000] "'male single'" "female div/dep/mar" "'male singl
e'" "'male single'" ...
##  $ OtherParties   : chr [1:1000] "none" "none" "none" "guarantor" ...
##  $ Residence      : int [1:1000] 4 2 3 4 4 4 4 2 4 2 ...
##  $ Asset          : chr [1:1000] "'real estate'" "'real estate'" "'real estate'"
"'life insurance'" ...
##  $ Age            : int [1:1000] 67 22 49 45 53 35 53 35 61 28 ...
##  $ OtherPlans     : chr [1:1000] "none" "none" "none" "none" ...
##  $ Housing        : chr [1:1000] "own" "own" "own" "'for free'" ...
##  $ ExistingCredits: int [1:1000] 2 1 1 1 2 1 1 1 1 2 ...
##  $ JobType        : chr [1:1000] "skilled" "skilled" "'unskilled resident'" "skill
ed" ...
##  $ Dependants     : int [1:1000] 1 1 2 2 2 2 1 1 1 1 ...
##  $ Phone          : chr [1:1000] "yes" "none" "none" "none" ...
##  $ Foreign        : chr [1:1000] "yes" "yes" "yes" "yes" ...
##  $ CreditRisk     : chr [1:1000] "good" "bad" "good" "good" ...
```

```
summarizeColumns(credit) %>% knitr::kable( caption = 'Table 1. Feature Summary before
Data Preprocessing')
```

Table 1. Feature Summary before Data Preprocessing

name	type	na	mean	disp	median	mad	min	max	nlevs

name	type	na	mean	disp	median	mad	min	max	nlevs
Checking	character	0	NA	0.6060000	NA	NA	63	394	4
Duration	integer	0	20.903	12.0588145	18.0	8.8956	4	72	0
CreditHistory	character	0	NA	0.4700000	NA	NA	40	530	5
Purpose	character	0	NA	0.7200000	NA	NA	9	280	10
CreditAmount	integer	0	3271.258	2822.7368760	2319.5	1627.1535	250	18424	0
Saving	character	0	NA	0.3970000	NA	NA	48	603	5
Employment	character	0	NA	0.6610000	NA	NA	62	339	5
Installment	integer	0	2.973	1.1187147	3.0	1.4826	1	4	0
Status	character	0	NA	0.4520000	NA	NA	50	548	4
OtherParties	character	0	NA	0.0930000	NA	NA	41	907	3
Residence	integer	0	2.845	1.1037179	3.0	1.4826	1	4	0
Asset	character	0	NA	0.6680000	NA	NA	154	332	4
Age	integer	0	35.546	11.3754686	33.0	10.3782	19	75	0
OtherPlans	character	0	NA	0.1860000	NA	NA	47	814	3
Housing	character	0	NA	0.2870000	NA	NA	108	713	3
ExistingCredits	integer	0	1.407	0.5776545	1.0	0.0000	1	4	0
JobType	character	0	NA	0.3700000	NA	NA	22	630	4
Dependants	integer	0	1.155	0.3620858	1.0	0.0000	1	2	0
Phone	character	0	NA	0.4040000	NA	NA	404	596	2
Foreign	character	0	NA	0.0370000	NA	NA	37	963	2
CreditRisk	character	0	NA	0.3000000	NA	NA	300	700	2

We next need to get rid of the excessive white space for all character variables.

```
credit[, sapply(credit, is.character )] <- sapply( credit[, sapply(credit, is.charact
er )], trimws)
```

We next check the levels for each character feature and the following points stood out a lot.
1. There were no visible missing values.
2. There were only 37 non foreign working individuals.
3. For Purpose we merged domestic appliance, furnture/equipment, repairs, radio/tv and retraining as household and new car and used car as Cars.
4. For Status, we saw that there was only 1 class for females so we renamed it female, and we merged the 3 classes of males to Male.
5. There were 200 unskilled individuals.
6. Only 49 people (4.9%) had all paid dues.
7. The maximum number of loans were for Radio/TV or for a new car.
8. Almost 60% of the individuals did not have a phone.

```
sapply( credit[ sapply(credit, is.character)], table)
```

```
## $Checking
##
##          '<0'       '>=200'    '0<=X<200' 'no checking'
##           274            63           269           394
##
## $CreditHistory
##
##                        'all paid' 'critical/other existing credit'
##                                49                              293
##             'delayed previously'                  'existing paid'
##                                88                              530
##            'no credits/all paid'
##                                40
##
## $Purpose
##
##            business  domestic appliance           education furniture/equipment
##                  97                  12                  50                 181
##             new car               other            radio/tv             repairs
##                 234                  12                 280                  22
##          retraining           used car
##                   9                 103
##
## $Saving
##
##             '<100'           '>=1000'       '100<=X<500'      '500<=X<1000'
##                603                 48                103                  63
## 'no known savings'
##                183
##
## $Employment
##
##       '<1'      '>=7'   '1<=X<4'   '4<=X<7' unemployed
##        172        253        339        174         62
##
## $Status
##
##     'male div/sep'     'male mar/wid'      'male single' female div/dep/mar
##                 50                 92                548                310
##
## $OtherParties
##
## 'co applicant'      guarantor           none
##             41             52            907
##
## $Asset
##
##    'life insurance' 'no known property'       'real estate'                 car
##                 232                 154                 282                 332
##
## $OtherPlans
##
##   bank   none stores
##    139    814     47
##
## $Housing
##
```

```
## 'for free'        own       rent
##        108        713        179
##
## $JobType
##
## 'high qualif/self emp/mgmt'   'unemp/unskilled non res'
##                         148                          22
##        'unskilled resident'                     skilled
##                         200                         630
##
## $Phone
##
## none  yes
##  596  404
##
## $Foreign
##
##  no yes
##  37 963
##
## $CreditRisk
##
##  bad good
##  300  700
```

We kept the priginal features intact to ensure data does not get affected and created new descriptive features from them. Each of the newly created feature had its name end with 1 to distinct itself from the original.

```
credit <- credit %>%
         mutate(Purpose1 = ifelse( Purpose %in% c('domestic appliance', 'furniture/eq
uipment', 'radio/tv', 'repairs',                                        'retraining'),
'Household',
                                   ifelse( Purpose %in% c('new car', 'used car'), 'Ca
rs',
                                           ifelse( grepl('business', Purpose), 'Busin
ess',
                                                   ifelse( grepl('education', Purpos
e), 'Education',
                                                           ifelse( grepl('other', Pur
pose), 'Other', Purpose))))),
                 Status1 = ifelse( Status != 'female div/dep/mar', 'Male', 'Female')
          )
```

We next changed all the character variables to factors.

```
credit[, sapply( credit, is.character )] <- lapply( credit[, sapply( credit, is.chara
cter )], factor)
```

Table 2 represents the summary statistics of the data after preprocesseing.

```
summarizeColumns(credit) %>% knitr::kable( caption = 'Table 2. Feature Summary after
 Data Preprocessing' )
```

Table 2. Feature Summary after Data Preprocessing

name	type	na	mean	disp	median	mad	min	max	nlevs

name	type	na	mean	disp	median	mad	min	max	nlevs
Checking	factor	0	NA	0.6060000	NA	NA	63	394	4
Duration	integer	0	20.903	12.0588145	18.0	8.8956	4	72	0
CreditHistory	factor	0	NA	0.4700000	NA	NA	40	530	5
Purpose	factor	0	NA	0.7200000	NA	NA	9	280	10
CreditAmount	integer	0	3271.258	2822.7368760	2319.5	1627.1535	250	18424	0
Saving	factor	0	NA	0.3970000	NA	NA	48	603	5
Employment	factor	0	NA	0.6610000	NA	NA	62	339	5
Installment	integer	0	2.973	1.1187147	3.0	1.4826	1	4	0
Status	factor	0	NA	0.4520000	NA	NA	50	548	4
OtherParties	factor	0	NA	0.0930000	NA	NA	41	907	3
Residence	integer	0	2.845	1.1037179	3.0	1.4826	1	4	0
Asset	factor	0	NA	0.6680000	NA	NA	154	332	4
Age	integer	0	35.546	11.3754686	33.0	10.3782	19	75	0
OtherPlans	factor	0	NA	0.1860000	NA	NA	47	814	3
Housing	factor	0	NA	0.2870000	NA	NA	108	713	3
ExistingCredits	integer	0	1.407	0.5776545	1.0	0.0000	1	4	0
JobType	factor	0	NA	0.3700000	NA	NA	22	630	4
Dependants	integer	0	1.155	0.3620858	1.0	0.0000	1	2	0
Phone	factor	0	NA	0.4040000	NA	NA	404	596	2
Foreign	factor	0	NA	0.0370000	NA	NA	37	963	2
CreditRisk	factor	0	NA	0.3000000	NA	NA	300	700	2
Purpose1	factor	0	NA	0.4960000	NA	NA	12	504	5
Status1	factor	0	NA	0.3100000	NA	NA	310	690	2

Data Exploration

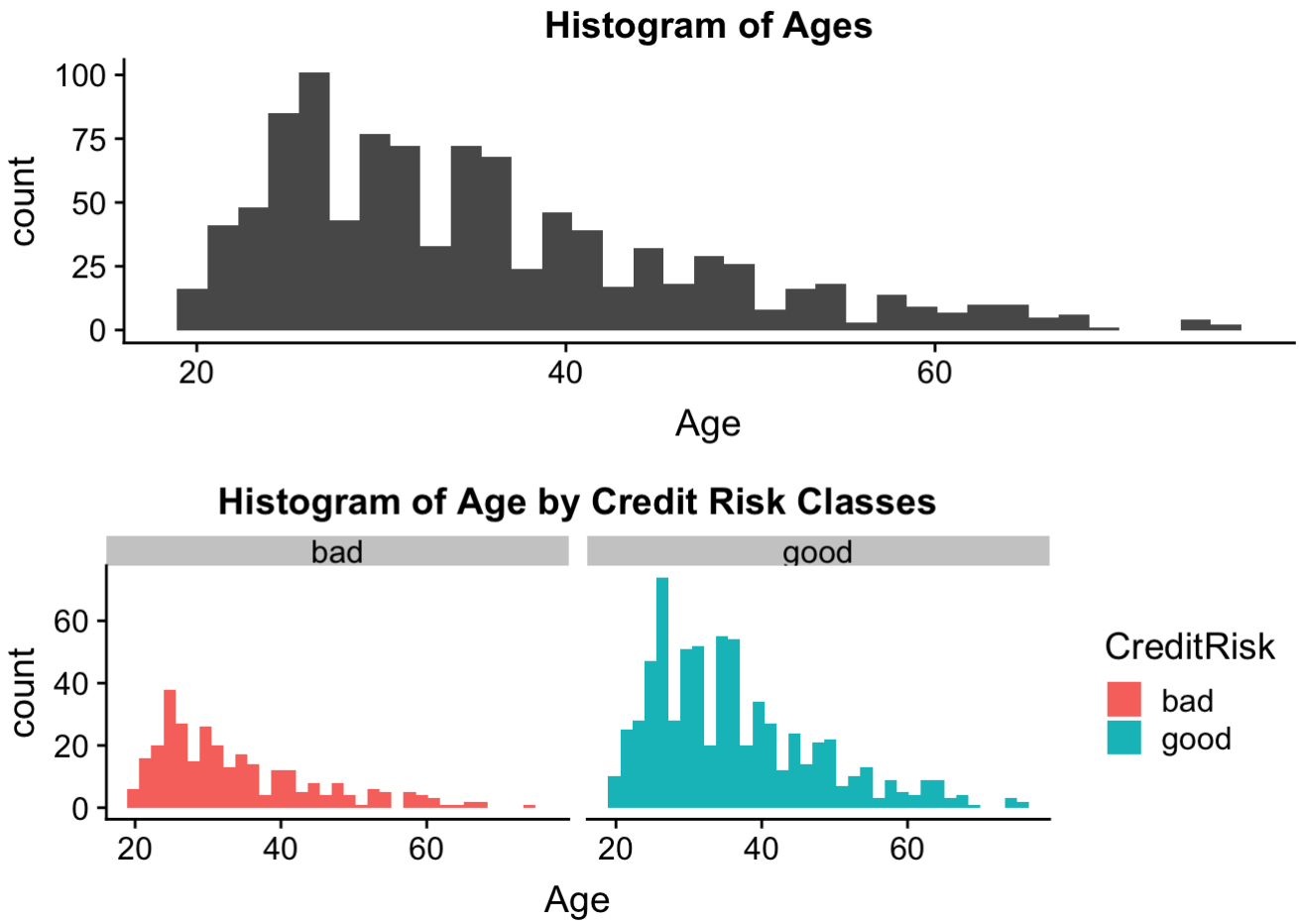
The next step in this stage of the project was to explore the data. For this, we first explored all the variables individually, followed by some multi variate relationships.

Univariate Visualisation

Numerical Features

Age

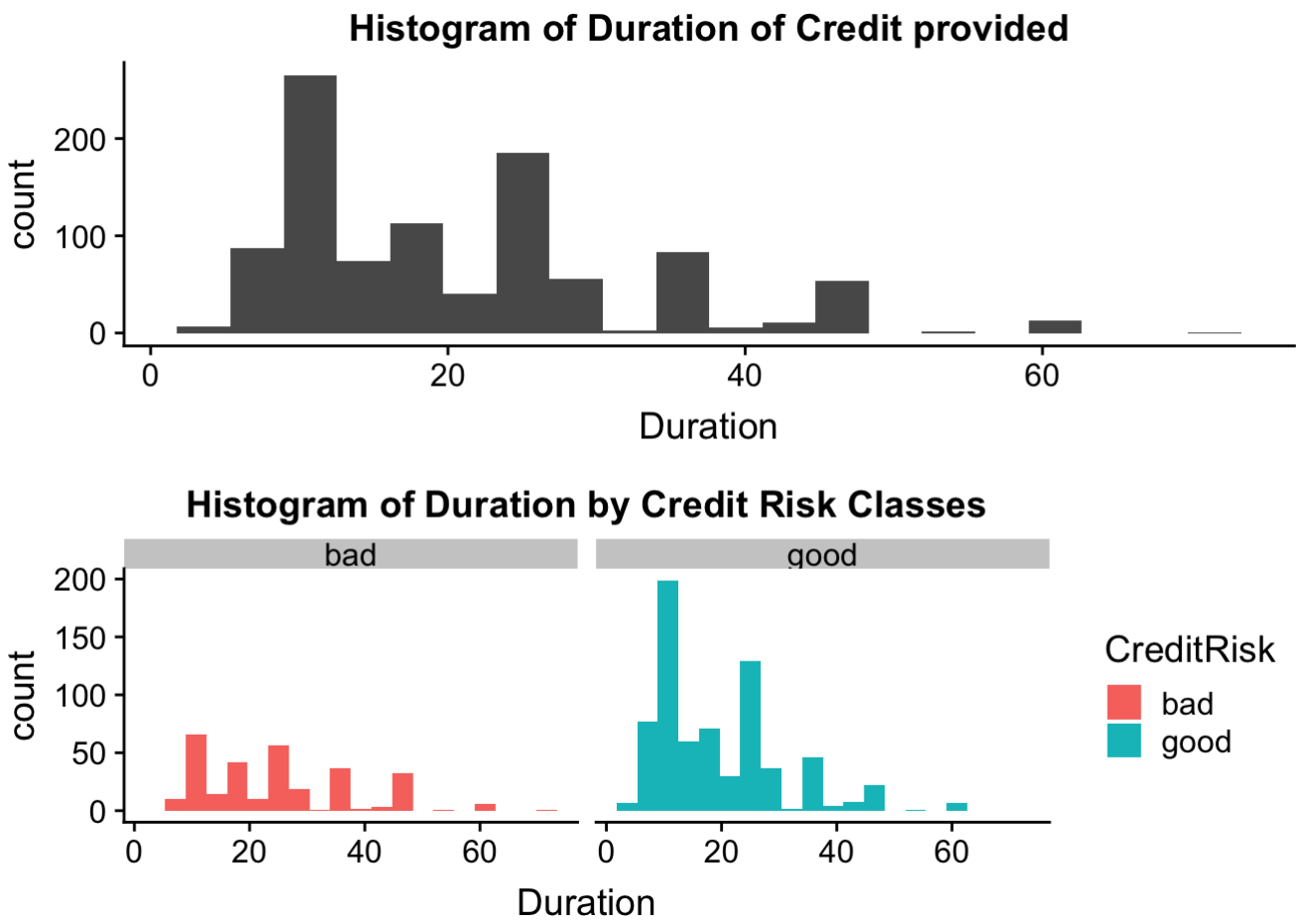
```
p1 <- ggplot(credit, aes(x = Age)) + geom_histogram(bins = 35) + labs(title = 'Histog
ram of Ages')
p2 <- ggplot(credit, aes(x = Age, fill = CreditRisk)) +
  geom_histogram(bins = 35) + facet_grid(~CreditRisk) +
  labs(title = 'Histogram of Age by Credit Risk Classes')
plot_grid(p1, p2, ncol = 1)
```



Most of individuals who would want to avail a credit facility are aged between 20 and 40. We see a normal distribution for credit risk. Individuals in this age would have commenced their carrier and hence would be able to take calculated risks.

Duration (in months)

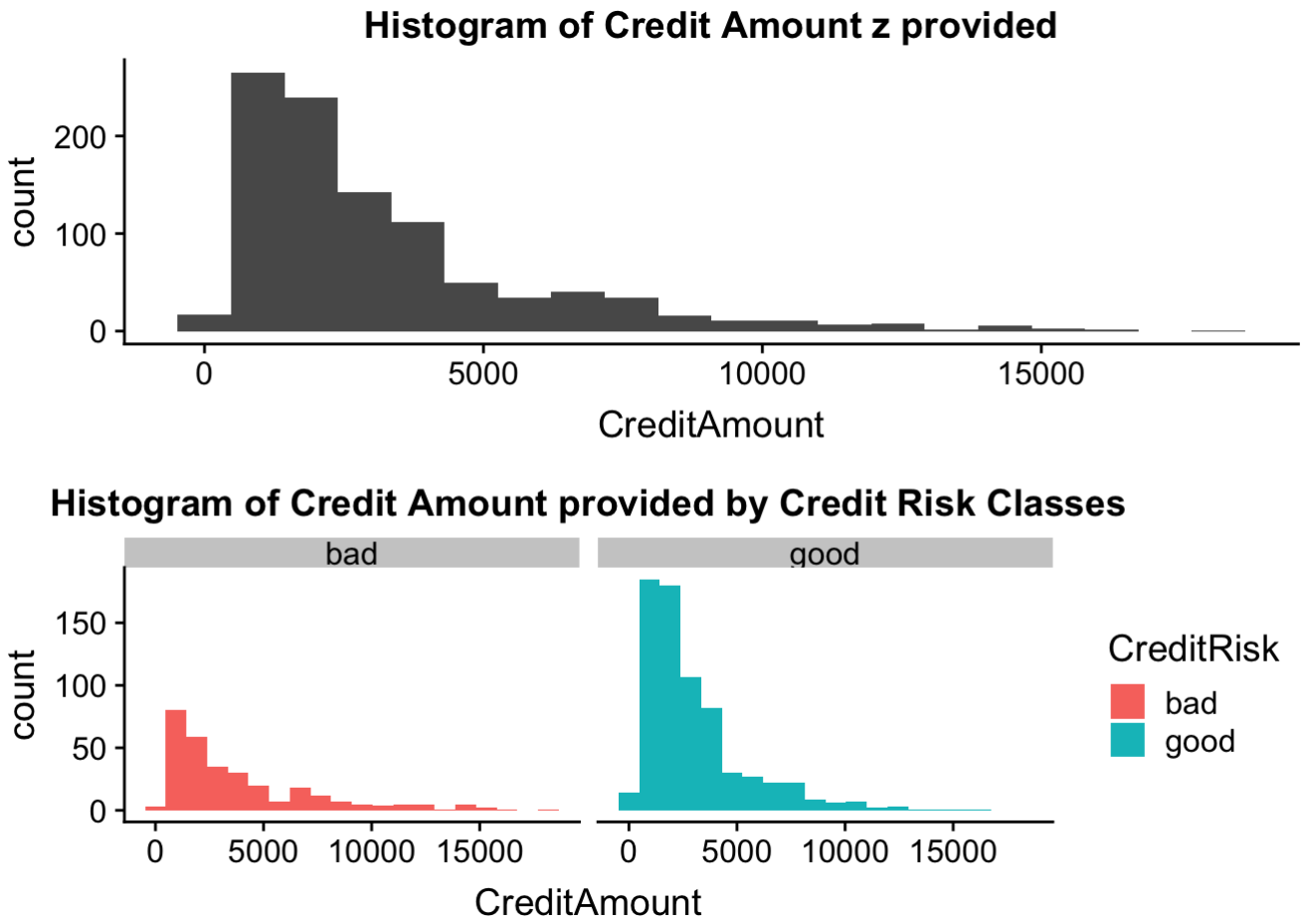
```
p3 <- ggplot(credit, aes(x = Duration)) + geom_histogram(bins = 20) +
  labs(title = 'Histogram of Duration of Credit provided')

p4 <- ggplot(credit, aes(x = Duration, fill = CreditRisk)) +
  geom_histogram(bins = 20) + facet_grid(~CreditRisk) +
  labs(title = 'Histogram of Duration by Credit Risk Classes')

plot_grid(p3, p4, ncol = 1)
```



Duration here refers to the tenure or length of period an individual agrees to pay out the credit facility. Durations appears to be right skewed with ideal tenure in the range of 12 to 24 months.

Credit Amount (in Deutsch Mark)

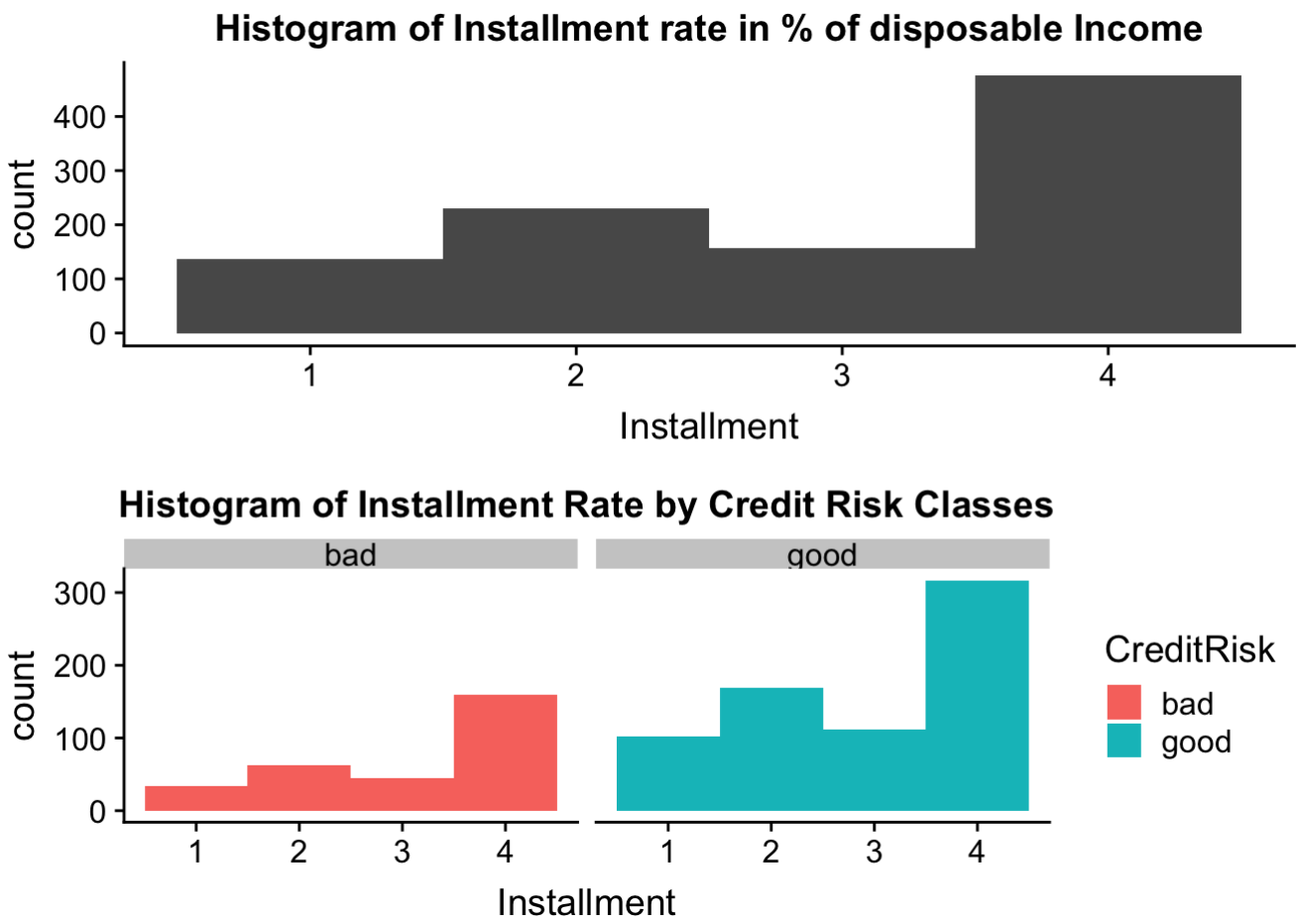
```
p5 <- ggplot(credit, aes(x = CreditAmount)) + geom_histogram(bins = 20) +
  labs(title = 'Histogram of Credit Amount z provided')

p6 <- ggplot(credit, aes(x = CreditAmount, fill = CreditRisk)) +
  geom_histogram(bins = 20) + facet_grid(~CreditRisk) +
  labs(title = 'Histogram of Credit Amount provided by Credit Risk Classes')

plot_grid(p5, p6, ncol = 1)
```



Credit amount is in Deutsch Mark. On aggregated level, individuals took a loan amount of approximately 1000 to 2000 marks, as indicated by the sharp kurtosis in the histogram of credit amount.

Installment (Rate in % of disposable income)

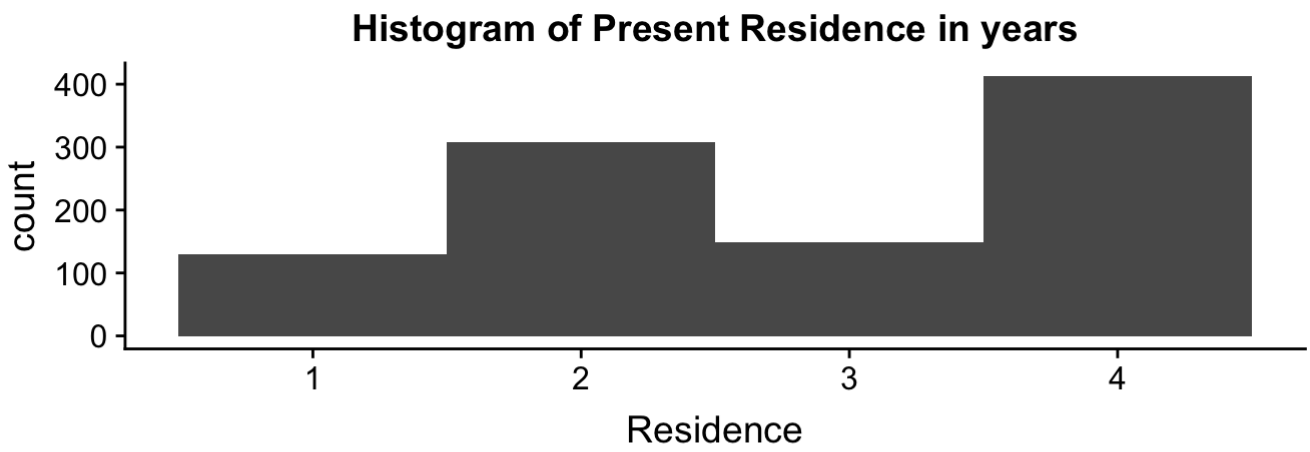
```
p7 <- ggplot(credit, aes(x = Installment)) + geom_histogram(bins = 4) +
  labs(title = 'Histogram of Installment rate in % of disposable Income')

p8 <- ggplot(credit, aes(x = Installment, fill = CreditRisk)) +
  geom_histogram(bins = 4) + facet_grid(~CreditRisk) +
  labs(title = 'Histogram of Installment Rate by Credit Risk Classes')

plot_grid(p7, p8, ncol = 1)
```



We are referring to the interest rate one would pay to avail the credit facility. Usually a lower instalment rate means, less interest payable. From the histogram, we can see that only good credit risk can avail a lower risk rate of 1-2 percent. Bad credit risk is right skewed as the majority instalment rate is aggregated to 4 percent.

Present Residence

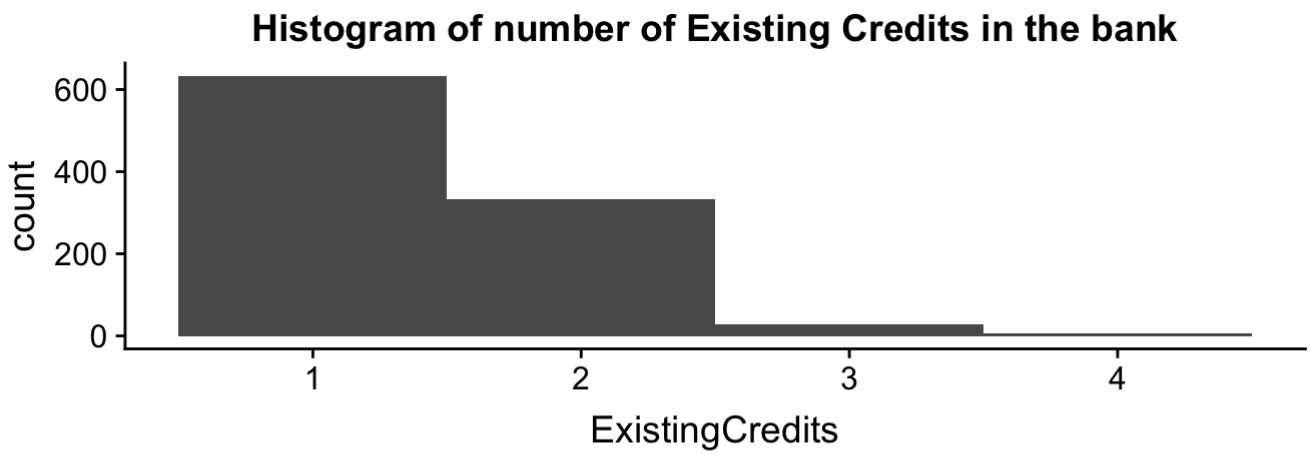
```
p9 <- ggplot(credit, aes(x = Residence)) + geom_histogram(bins = 4) +
  labs(title = 'Histogram of Present Residence in years')

p10 <- ggplot(credit, aes(x = Residence, fill = CreditRisk)) +
  geom_histogram(bins = 4) + facet_grid(~CreditRisk) +
  labs(title = 'Histogram of Present Residence by Credit Risk Classes')

plot_grid(p9, p10, ncol = 1)
```

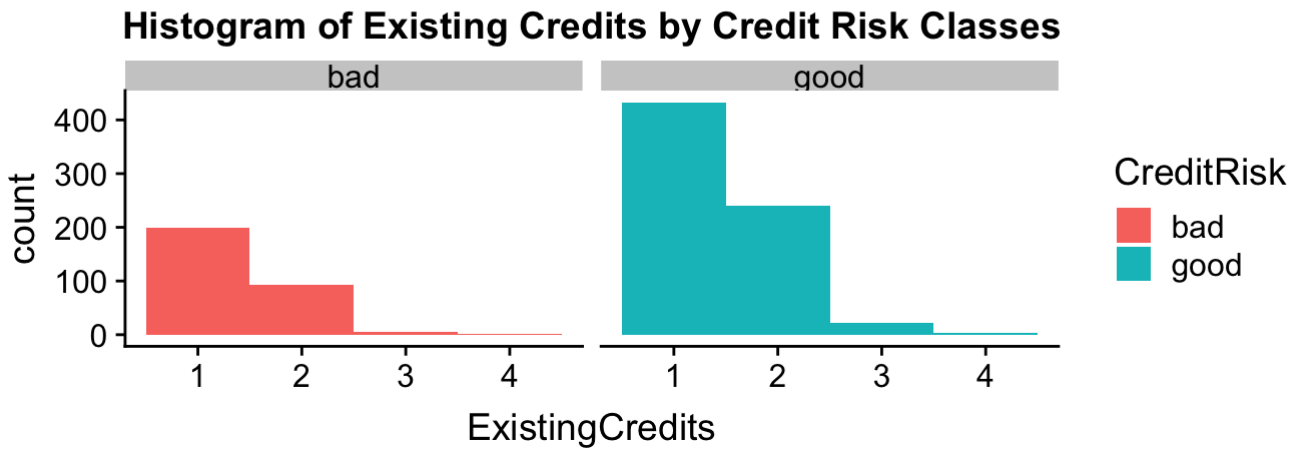


Number of years an individual resides in the current address, determines their residential status and shows stability of individual movement. For credit risk a long stay in the current address determines a lower risk and better risk-taking ability. Hence the histogram is right skewed for good credit risk.

Existing Credits

```
p11 <- ggplot(credit, aes(x = ExistingCredits)) + geom_histogram(bins = 4) +
  labs(title = 'Histogram of number of Existing Credits in the bank')

p12 <- ggplot(credit, aes(x = ExistingCredits, fill = CreditRisk)) +
  geom_histogram(bins = 4) + facet_grid(~CreditRisk) +
  labs(title = 'Histogram of Existing Credits by Credit Risk Classes')

plot_grid(p11, p12, ncol = 1)
```

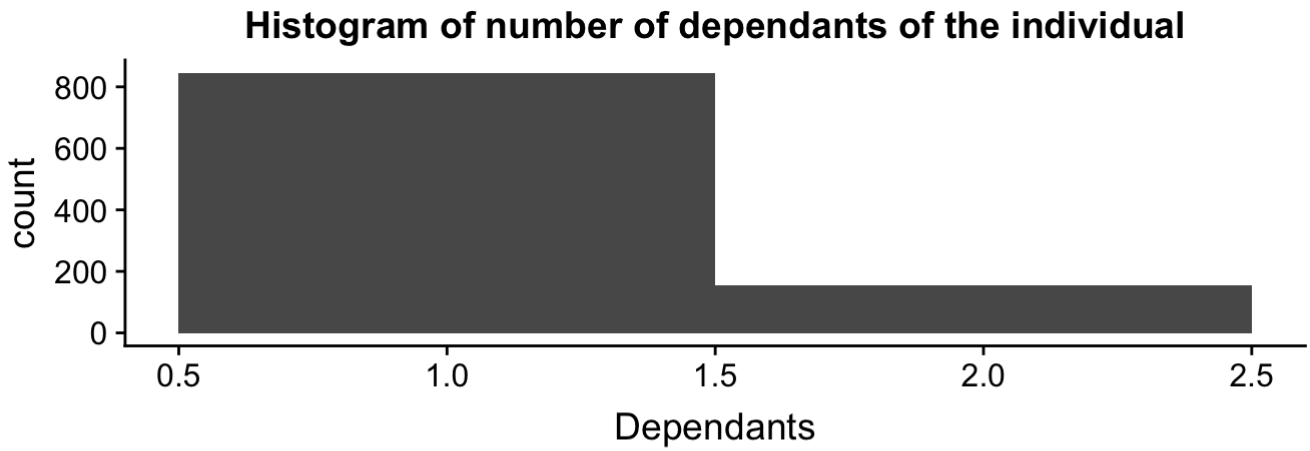


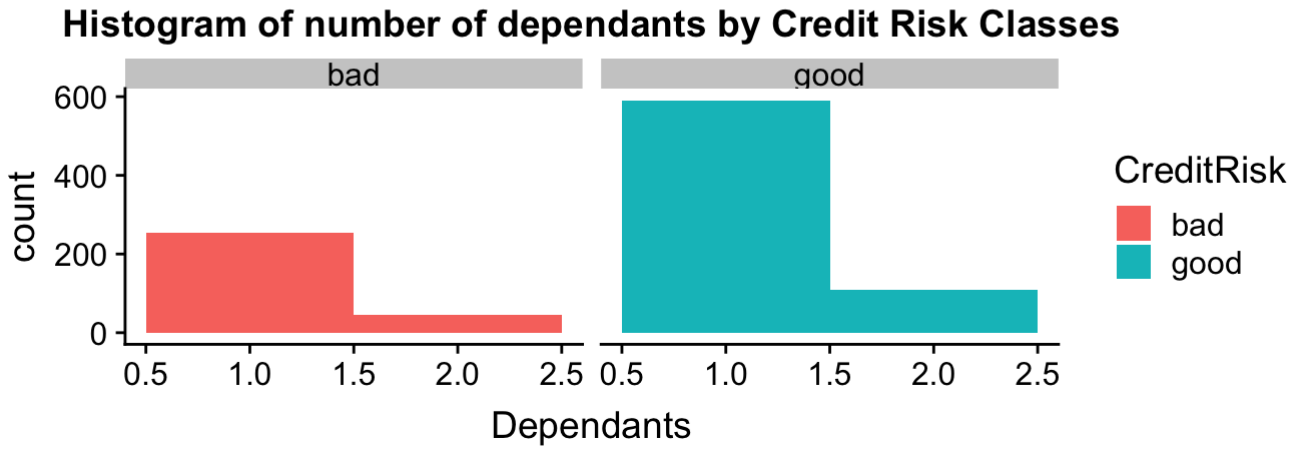


Number of existing credit facilities at a bank plays an important role in determining credit risk. Existing debt also means the capability to take more debt decreases. Hence, number of existing credit facilities limited to 1-2. Higher number of existing credit are more likely to be good credit risk.

Number of Dependants

```
p13 <- ggplot(credit, aes(x = Dependants)) + geom_histogram(bins = 2) +
  labs(title = 'Histogram of number of dependants of the individual')

p14 <- ggplot(credit, aes(x = Dependants, fill = CreditRisk)) +
  geom_histogram(bins = 2) + facet_grid(~CreditRisk) +
  labs(title = 'Histogram of number of dependants by Credit Risk Classes')

plot_grid(p13, p14, ncol = 1)
```



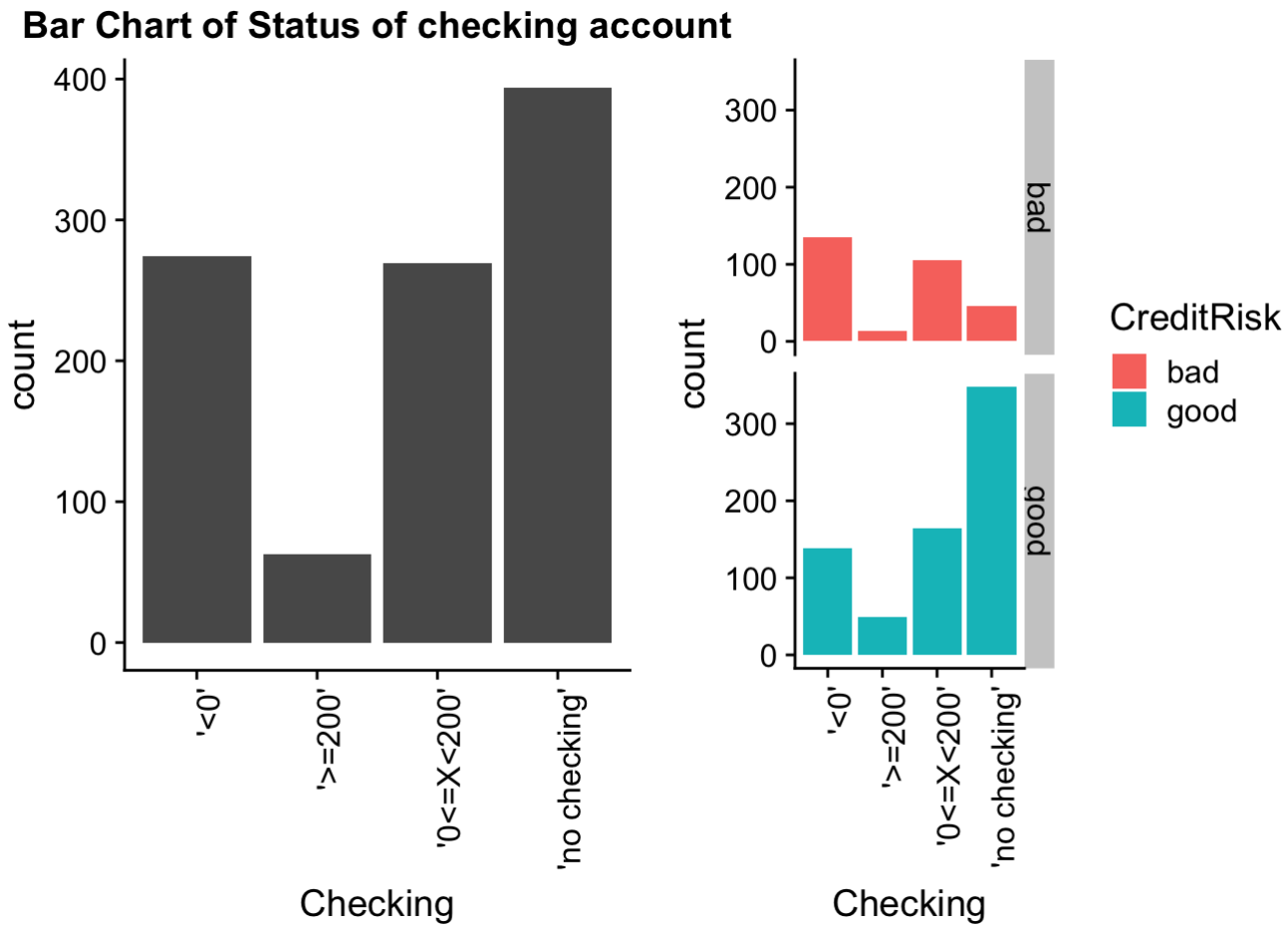


This factor establishes living cost of an individual. People with higher number dependents to have higher living cost when compared to no dependents. As the histogram is right skewed, it suggests that, individual with no dependents or less dependents are more likely to borrow more credit.

Categorical Features

Status of checking account in bank

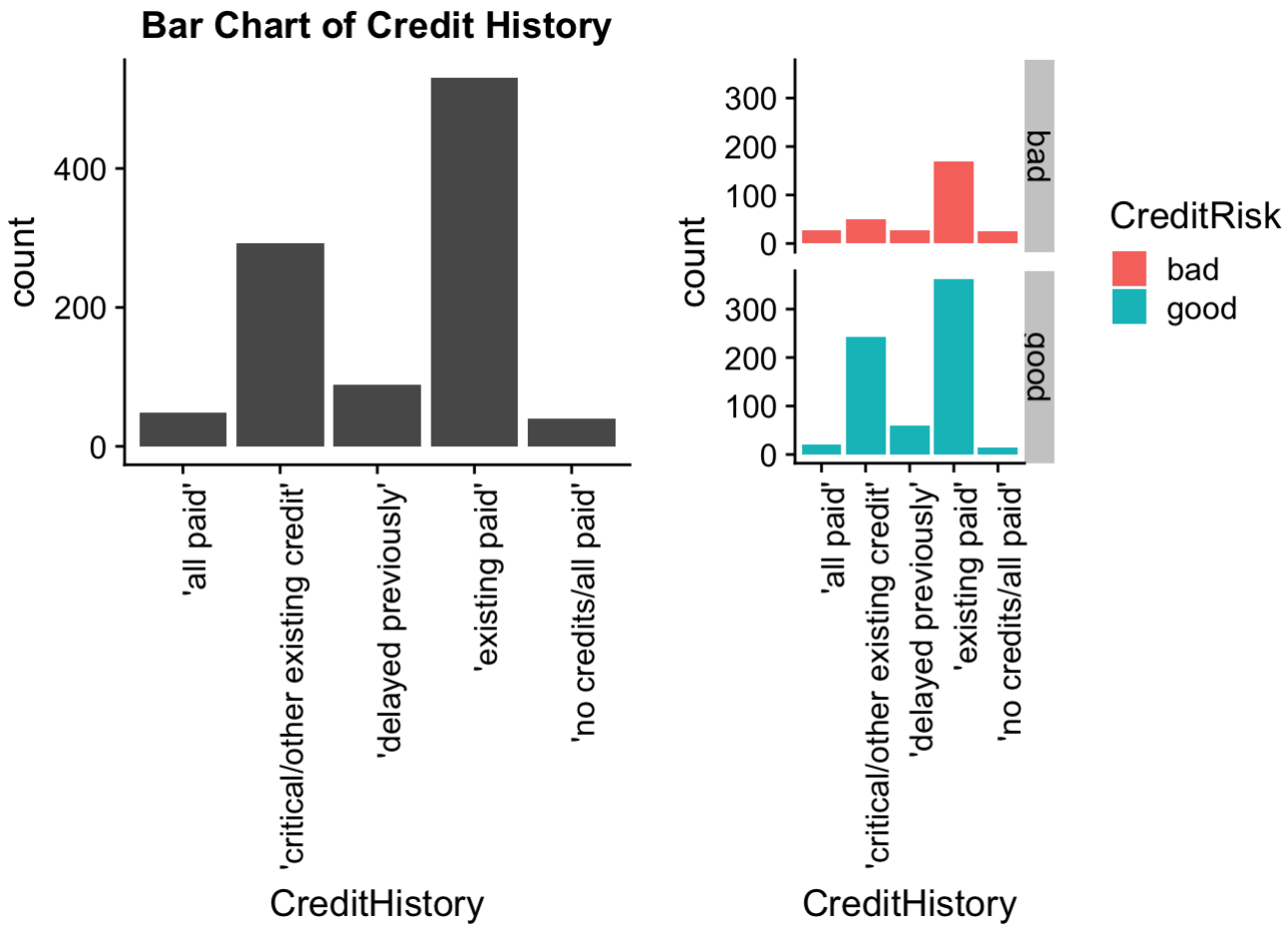
```
p15 <- ggplot(credit, aes(x = Checking)) + geom_bar() +
  labs(title = 'Bar Chart of Status of checking account') +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

p16 <- ggplot(credit, aes(x = Checking, fill = CreditRisk)) +
  geom_bar() + facet_grid(CreditRisk~.) +
  labs(title = '') +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

plot_grid(p15, p16, ncol = 2)
```



Number of people not having a checking account with the bank was considerably high. Individuals with checking account facility of less than 200 were more likely to avail credit facilities and Individuals with checking account facility of less than 0 were likely to be grouped in bad credit risk

Credit History

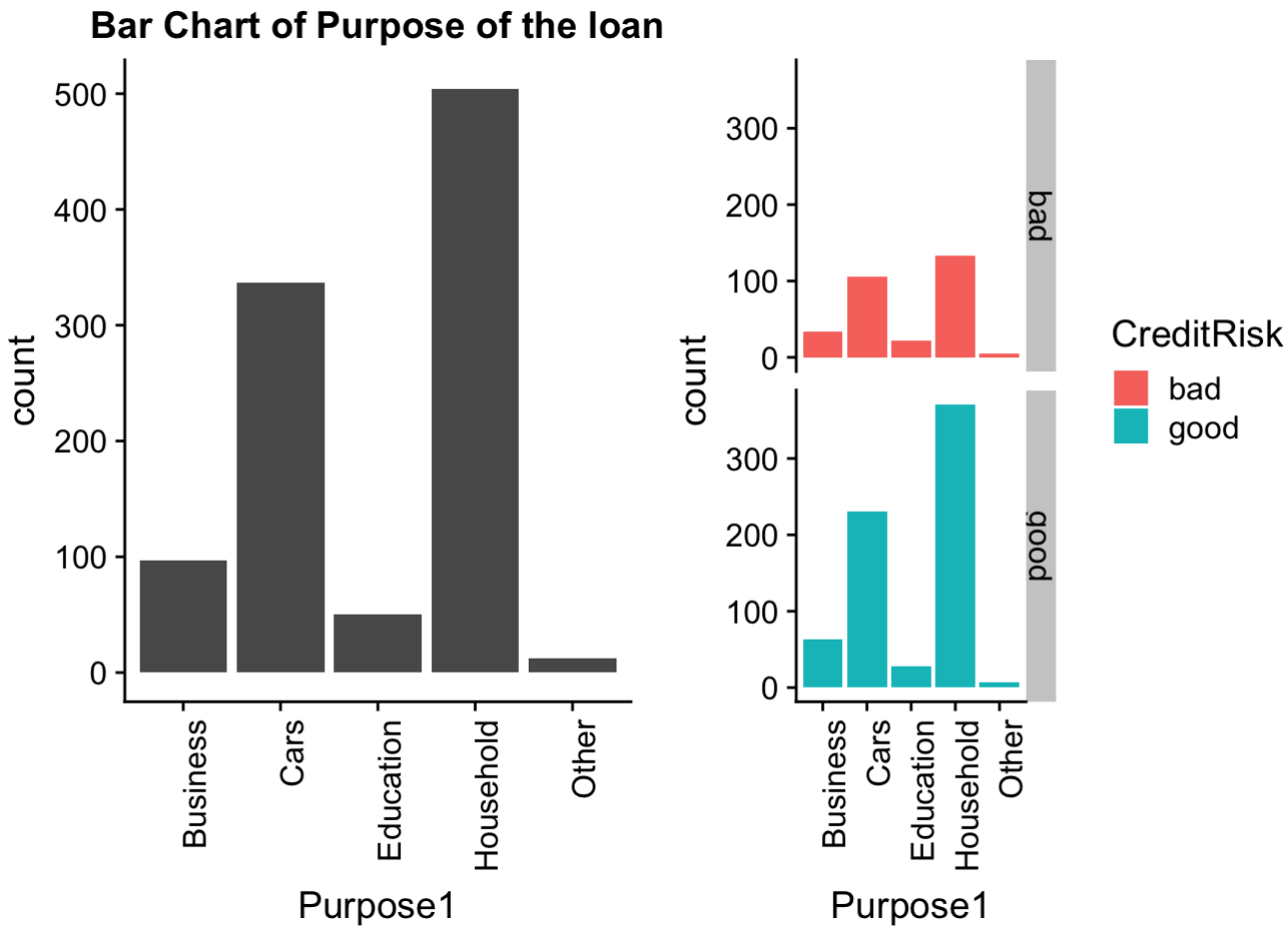
```
p17 <- ggplot(credit, aes(x = CreditHistory)) + geom_bar() +
  labs(title = 'Bar Chart of Credit History') +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

p18 <- ggplot(credit, aes(x = CreditHistory, fill = CreditRisk)) +
  geom_bar() + facet_grid(CreditRisk~.) +
  labs(title = '') +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

plot_grid(p17, p18, ncol = 2)
```



When making lending decisions, lenders review your credit history to determine how likely an individual would repay your loan on time. A longer history shows you have more experience using credit and hence lenders can be more accurate in determining the level of risk they take on when lending to you.

Purpose

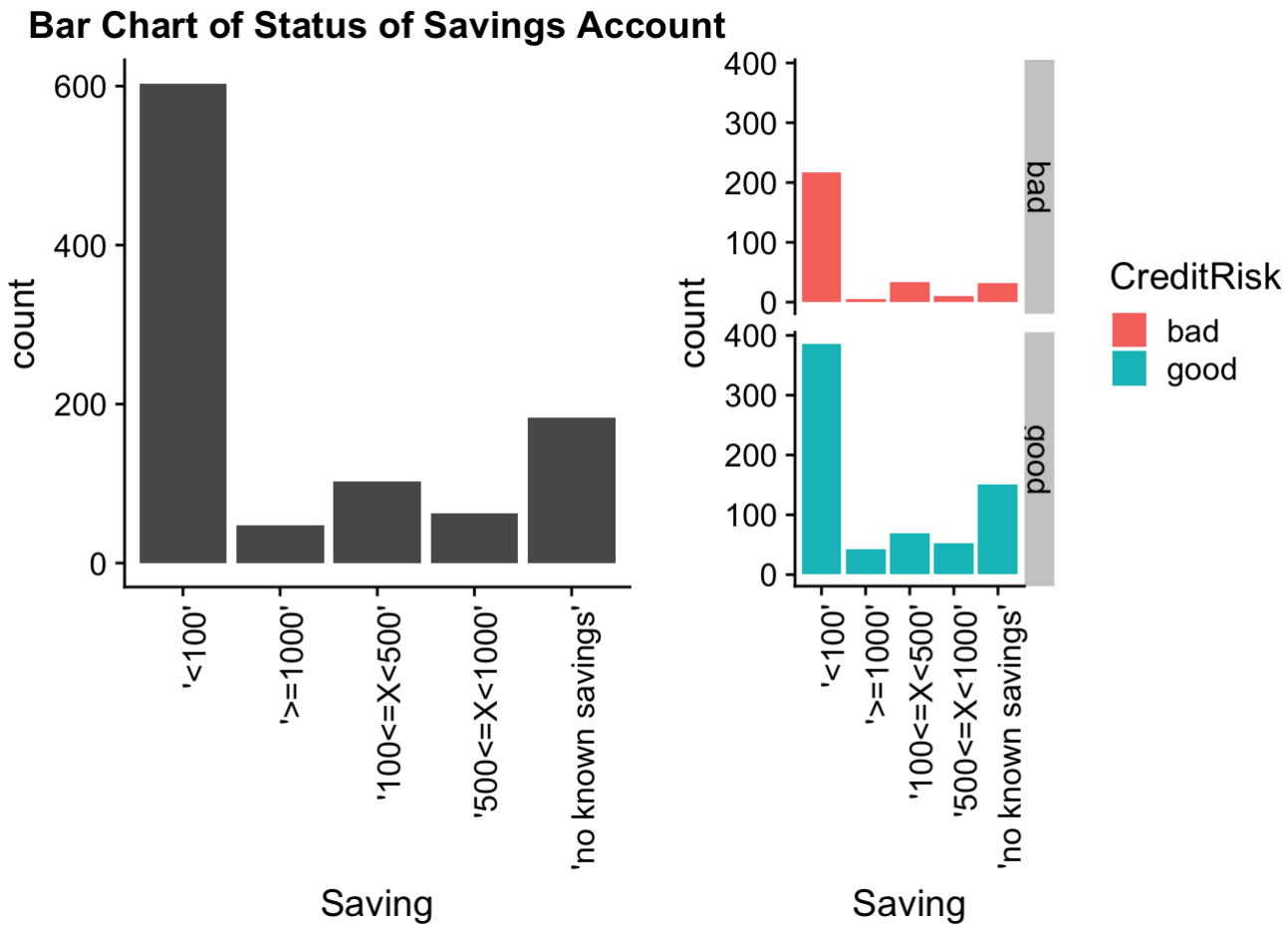
```
p19 <- ggplot(credit, aes(x = Purpose1)) + geom_bar() +
  labs(title = 'Bar Chart of Purpose of the loan') +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

p20 <- ggplot(credit, aes(x = Purpose1, fill = CreditRisk)) +
  geom_bar() + facet_grid(CreditRisk~.) +
  labs(title = '') +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

plot_grid(p19, p20, ncol = 2)
```



Loan purpose has been divides into four categories. Majority of people usually take a credit facility to purchase a car or make house hold improvements.

Status of Savings Account

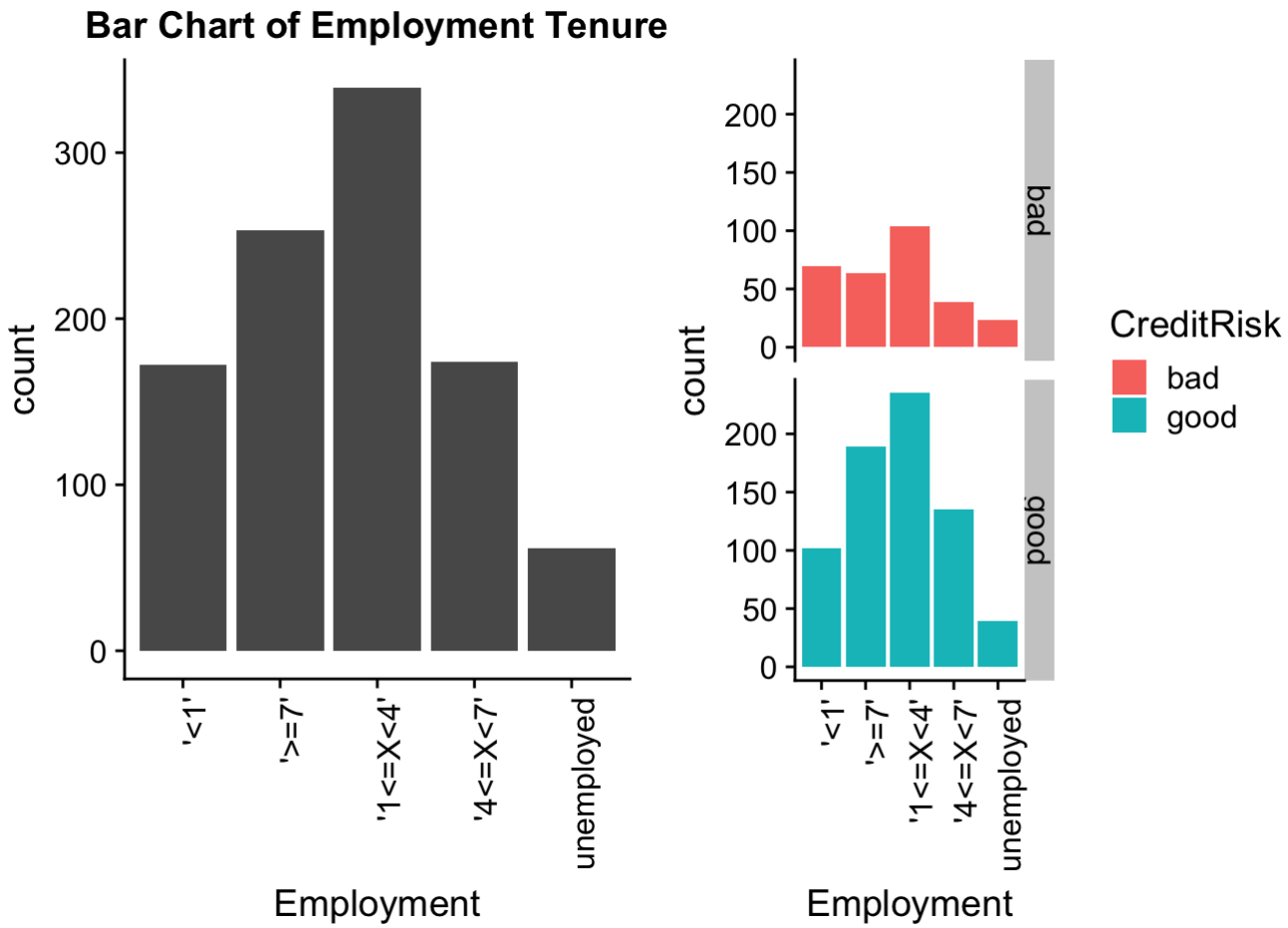
```
p21 <- ggplot(credit, aes(x = Saving)) + geom_bar() +
  labs(title = 'Bar Chart of Status of Savings Account') +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

p22 <- ggplot(credit, aes(x = Saving, fill = CreditRisk)) +
  geom_bar() + facet_grid(CreditRisk~.) +
  labs(title = '') +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

plot_grid(p21, p22, ncol = 2)
```



Savings or cash held in once savings account are divided into four categories. Savings of less than 100 thousand marks are likely to borrow credit facility. Credit risk remains good when individual have more than 100 thousand marks in savings.

Employment

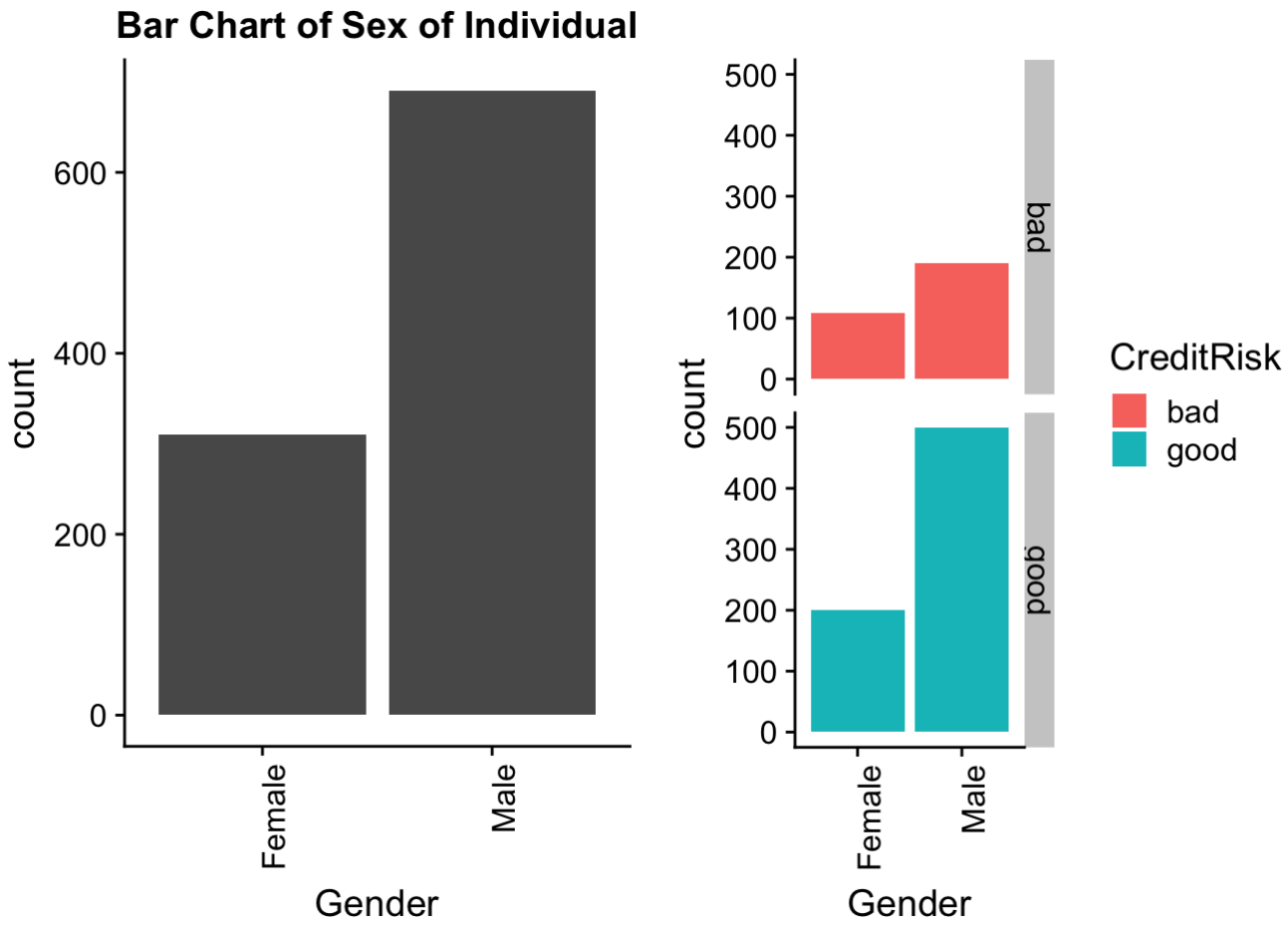
```
p23 <- ggplot(credit, aes(x = Employment)) + geom_bar() +
  labs(title = 'Bar Chart of Employment Tenure') +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

p24 <- ggplot(credit, aes(x = Employment, fill = CreditRisk)) +
  geom_bar() + facet_grid(CreditRisk~.) +
  labs(title = '') +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

plot_grid(p23, p24, ncol = 2)
```



Segregating employment tenure by credit risk showed that good credit risk individuals were more likely to have higher employment tenure whereas bed credit risk individuals were mostly lower employment tenure.

Sex of Individual

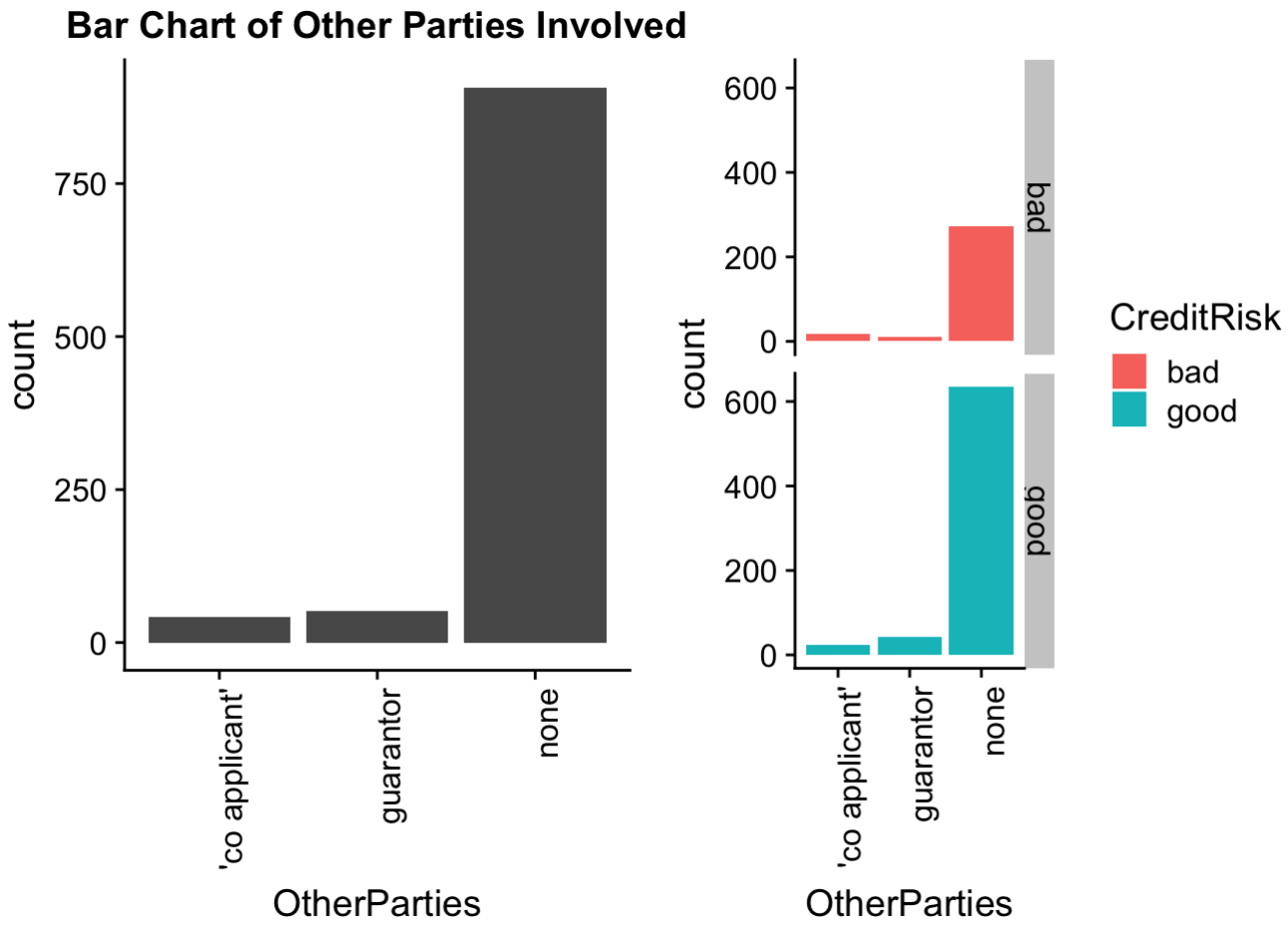
```
p25 <- ggplot(credit, aes(x = Status1)) + geom_bar() +
  labs(title = 'Bar Chart of Sex of Individual', x = "Gender") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

p26 <- ggplot(credit, aes(x = Status1, fill = CreditRisk)) +
  geom_bar() + facet_grid(CreditRisk~.) +
  labs(title = '', x = "Gender") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

plot_grid(p25, p26, ncol = 2)
```



Males dominated the overall credit risk taking ability. However, the bad credit risk was proportionality more for the females.

Other Parties Involved

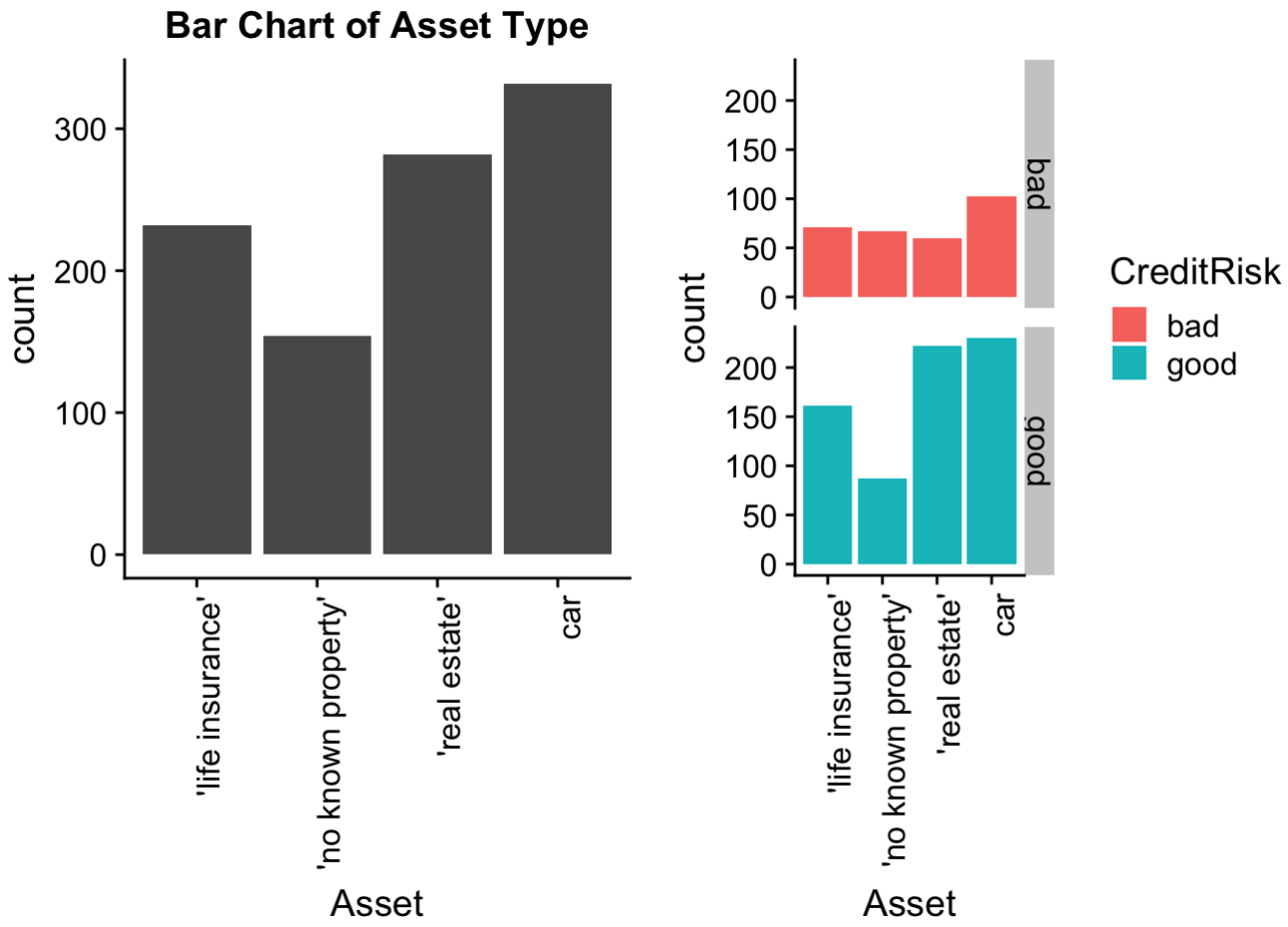
```
p27 <- ggplot(credit, aes(x = OtherParties)) + geom_bar() +
  labs(title = 'Bar Chart of Other Parties Involved') +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

p28 <- ggplot(credit, aes(x = OtherParties, fill = CreditRisk)) +
  geom_bar() + facet_grid(CreditRisk~.) +
  labs(title = '') +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

plot_grid(p27, p28, ncol = 2)
```



In total three parties were involved – co applicant, guarantor and none. None stood out among the three, as it also included individuals who had applied for credit facility alone.

Asset Type

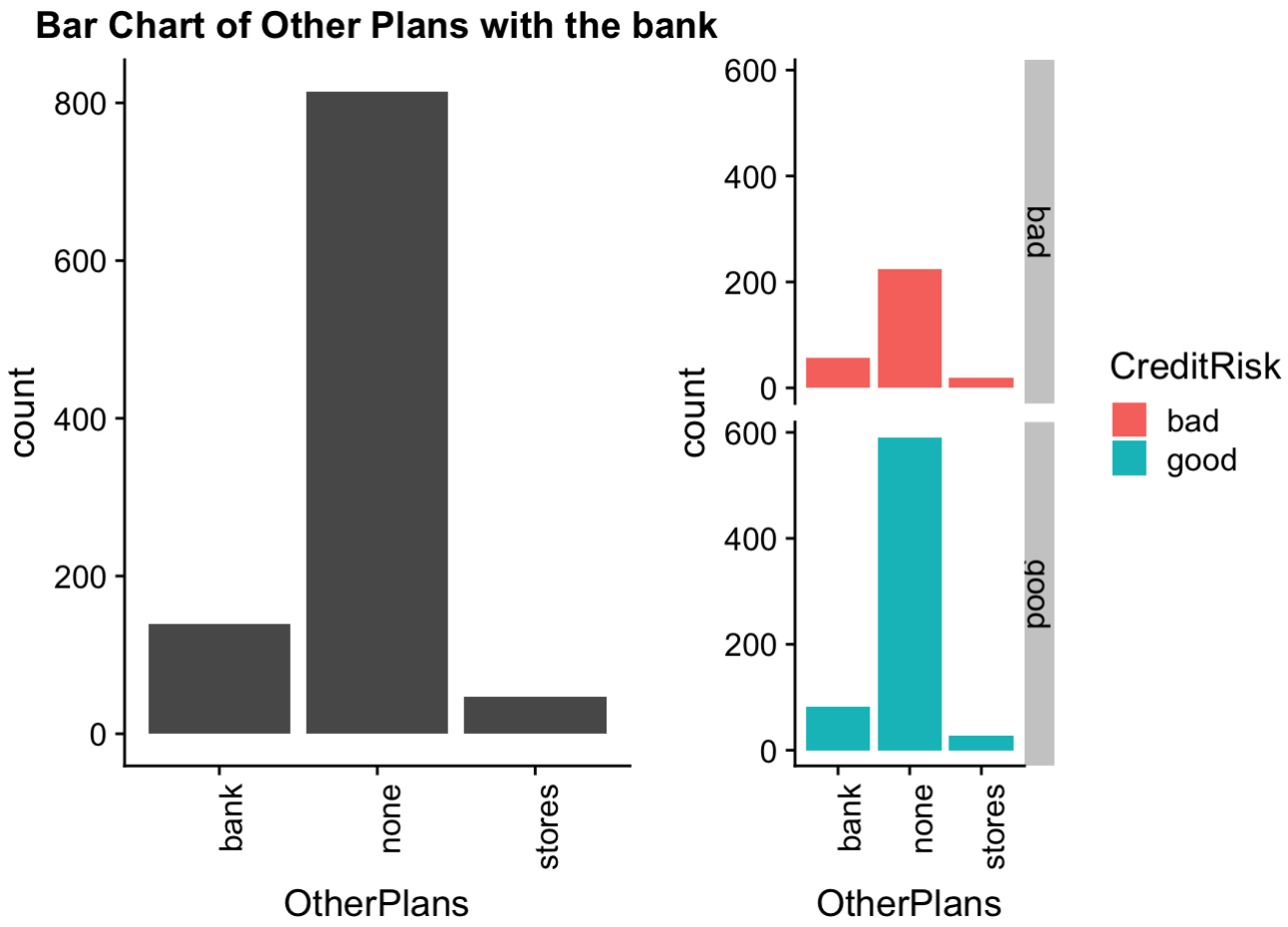
```
p29 <- ggplot(credit, aes(x = Asset)) + geom_bar() +
  labs(title = 'Bar Chart of Asset Type') +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

p30 <- ggplot(credit, aes(x = Asset, fill = CreditRisk)) +
  geom_bar() + facet_grid(CreditRisk~.) +
  labs(title = '') +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

plot_grid(p29, p30, ncol = 2)
```



Individuals are more likely to be grouped in good credit risk if one has listed a real estate or car as an asset. On the other hand, people who do not own a real estate are more likely to default or fall in bad credit risk. One of the reason behind this could be that, people who own a piece of real estate property are more likely to avail credit facility at a lower interest, as it would fall under secured lending.

Other Plans with the bank

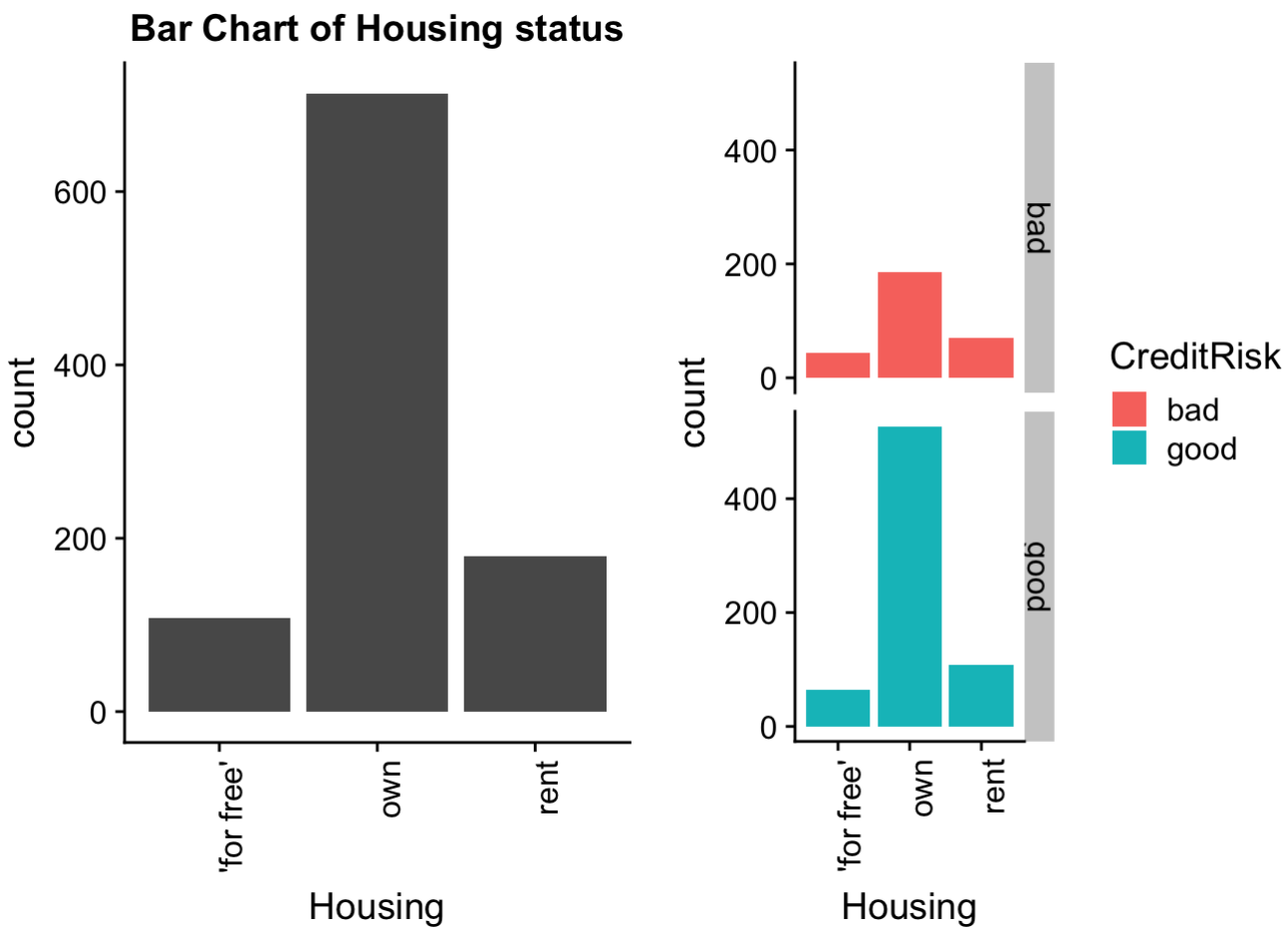
```
p31 <- ggplot(credit, aes(x = OtherPlans)) + geom_bar() +
  labs(title = 'Bar Chart of Other Plans with the bank') +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

p32 <- ggplot(credit, aes(x = OtherPlans, fill = CreditRisk)) +
  geom_bar() + facet_grid(CreditRisk~.) +
  labs(title = '') +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

plot_grid(p31, p32, ncol = 2)
```



Individuals who do not have any other commitments towards other bank, stores or other facility would have more appetite for credit facility, as their undistributed income would be higher.

Housing

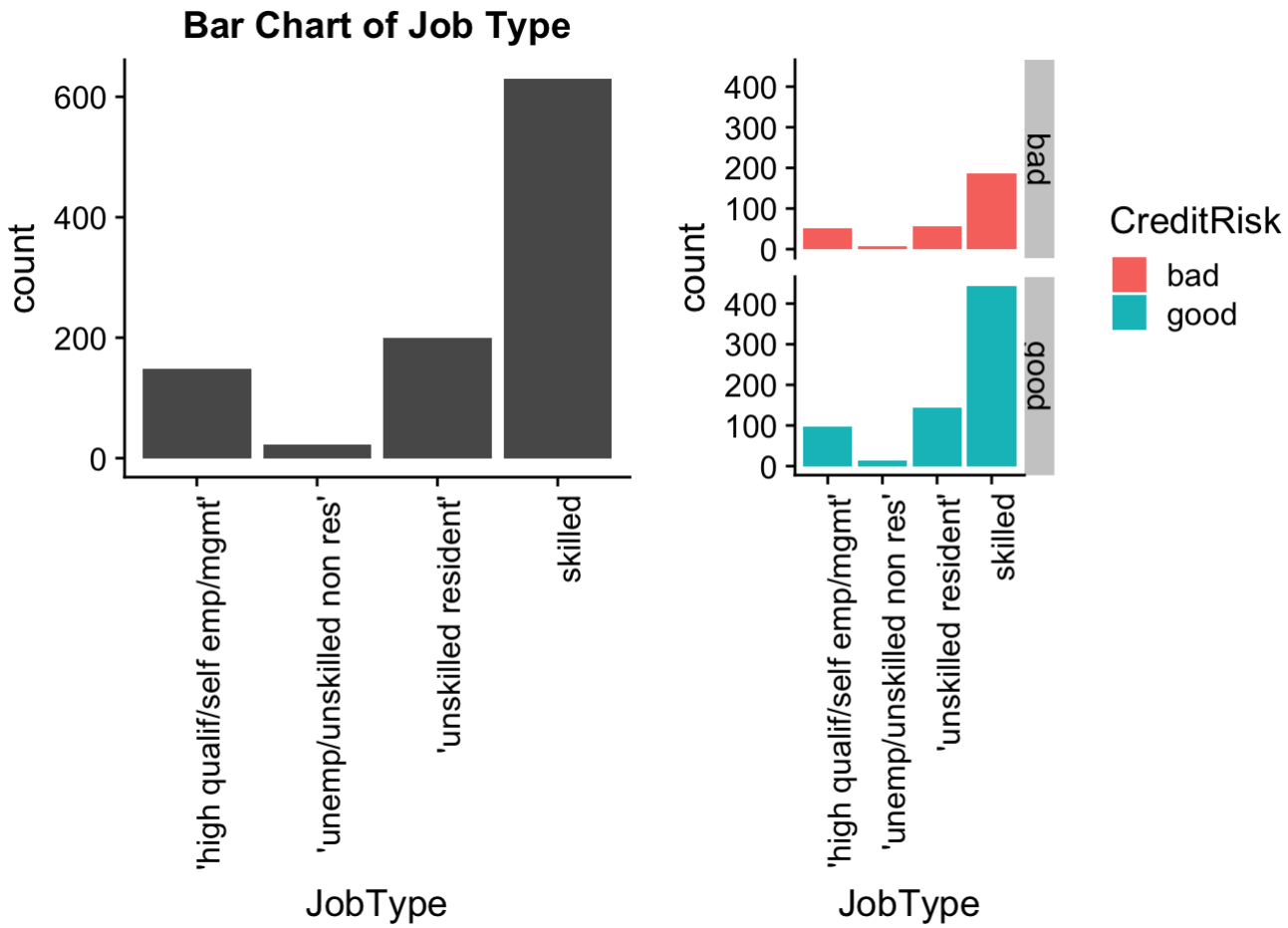
```
p33 <- ggplot(credit, aes(x = Housing)) + geom_bar() +
  labs(title = 'Bar Chart of Housing status') +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

p34 <- ggplot(credit, aes(x = Housing, fill = CreditRisk)) +
  geom_bar() + facet_grid(CreditRisk~.) +
  labs(title = '') +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

plot_grid(p33, p34, ncol = 2)
```



Most individuals own a house. Possibility to falling into bad credit risk group increases when living arrangement status is rent. People living for free do not have any form of commitment in terms of living expense and hence possibility of falling into bad credit risk is minimal.

Job Type

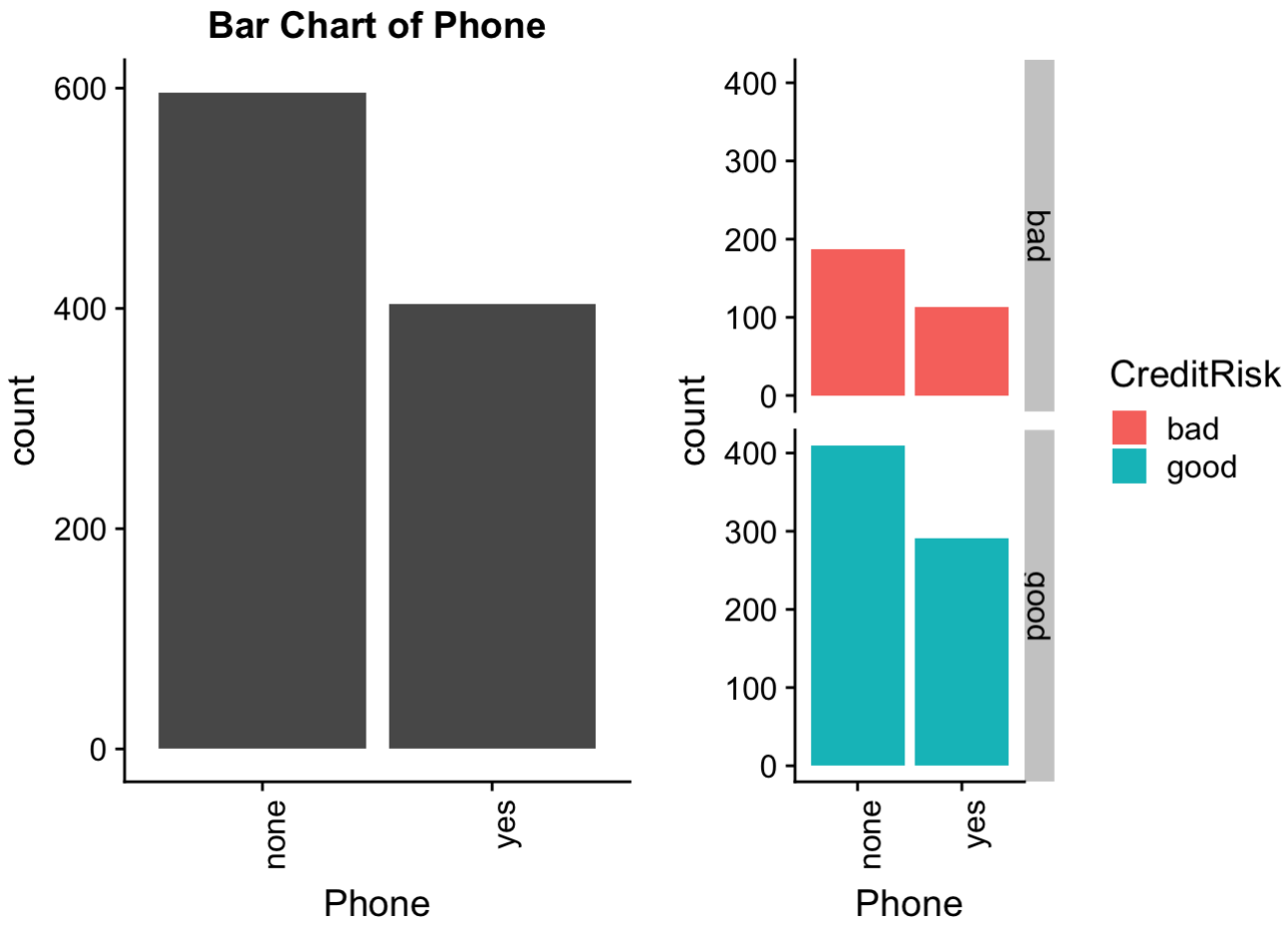
```
p35 <- ggplot(credit, aes(x = JobType)) + geom_bar() +
  labs(title = 'Bar Chart of Job Type') +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

p36 <- ggplot(credit, aes(x = JobType, fill = CreditRisk)) +
  geom_bar() + facet_grid(CreditRisk~.) +
  labs(title = '') +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

plot_grid(p35, p36, ncol = 2)
```



If unemployed, possibility of getting credit facility remains minimal. Skilled job type individual availed the maximum credit risk facility. Possibility of an unskilled job type to fall into bad credit risk is proportionality higher.

Phone

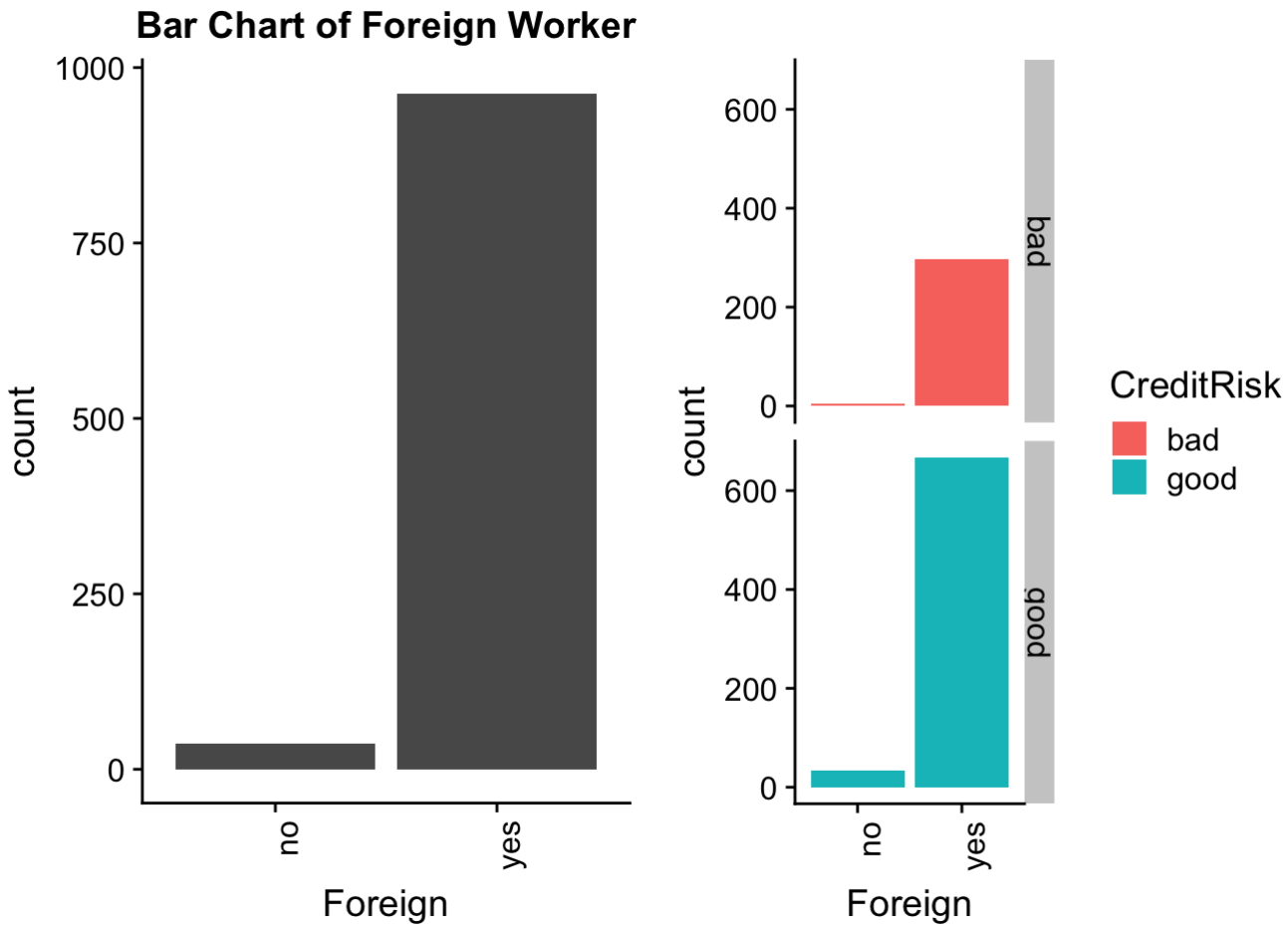
```
p37 <- ggplot(credit, aes(x = Phone)) + geom_bar() +
  labs(title = 'Bar Chart of Phone') +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

p38 <- ggplot(credit, aes(x = Phone, fill = CreditRisk)) +
  geom_bar() + facet_grid(CreditRisk~.) +
  labs(title = '') +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

plot_grid(p37, p38, ncol = 2)
```



Most individuals do not have a phone connection. However, this factor did not play a significant role in availing credit faculty.

Foreign Worker

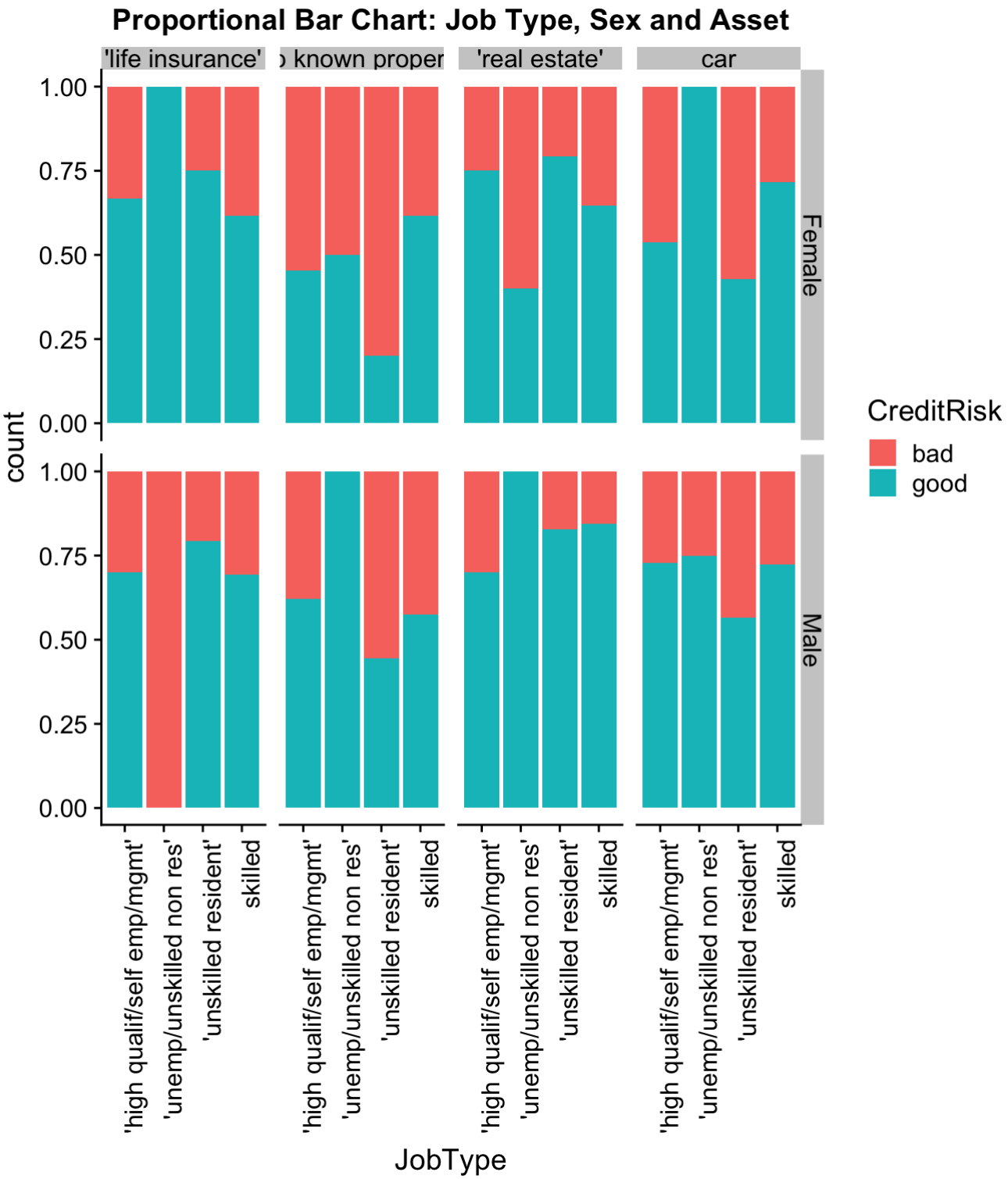
```
p39 <- ggplot(credit, aes(x = Foreign)) + geom_bar() +
  labs(title = 'Bar Chart of Foreign Worker') +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

p40 <- ggplot(credit, aes(x = Foreign, fill = CreditRisk)) +
  geom_bar() + facet_grid(CreditRisk~.) +
  labs(title = '') +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

plot_grid(p39, p40, ncol = 2)
```



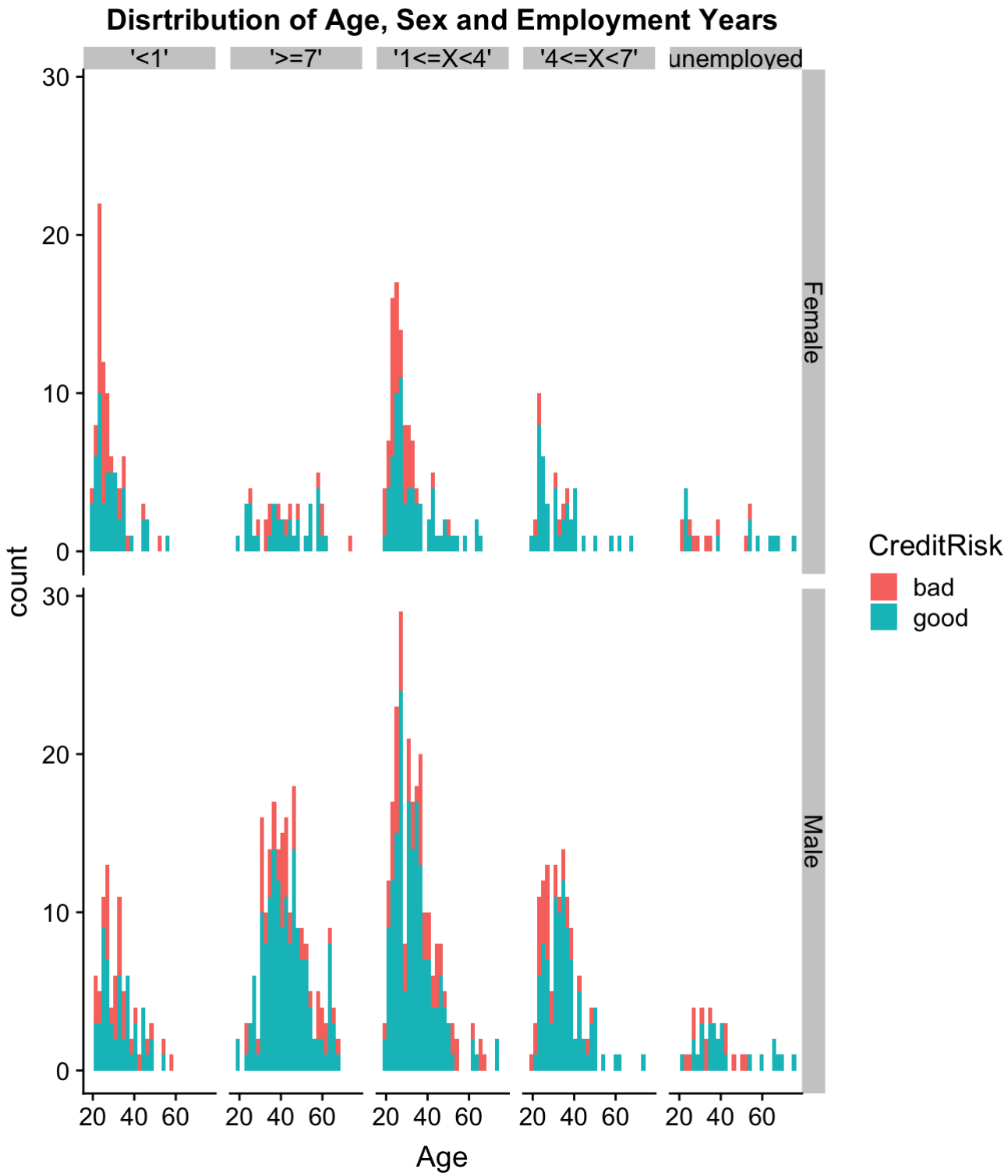
Nationality played a significant role in availing credit facility. Most individuals who availed credit facility where foreigners.

Multivariate Visualisation

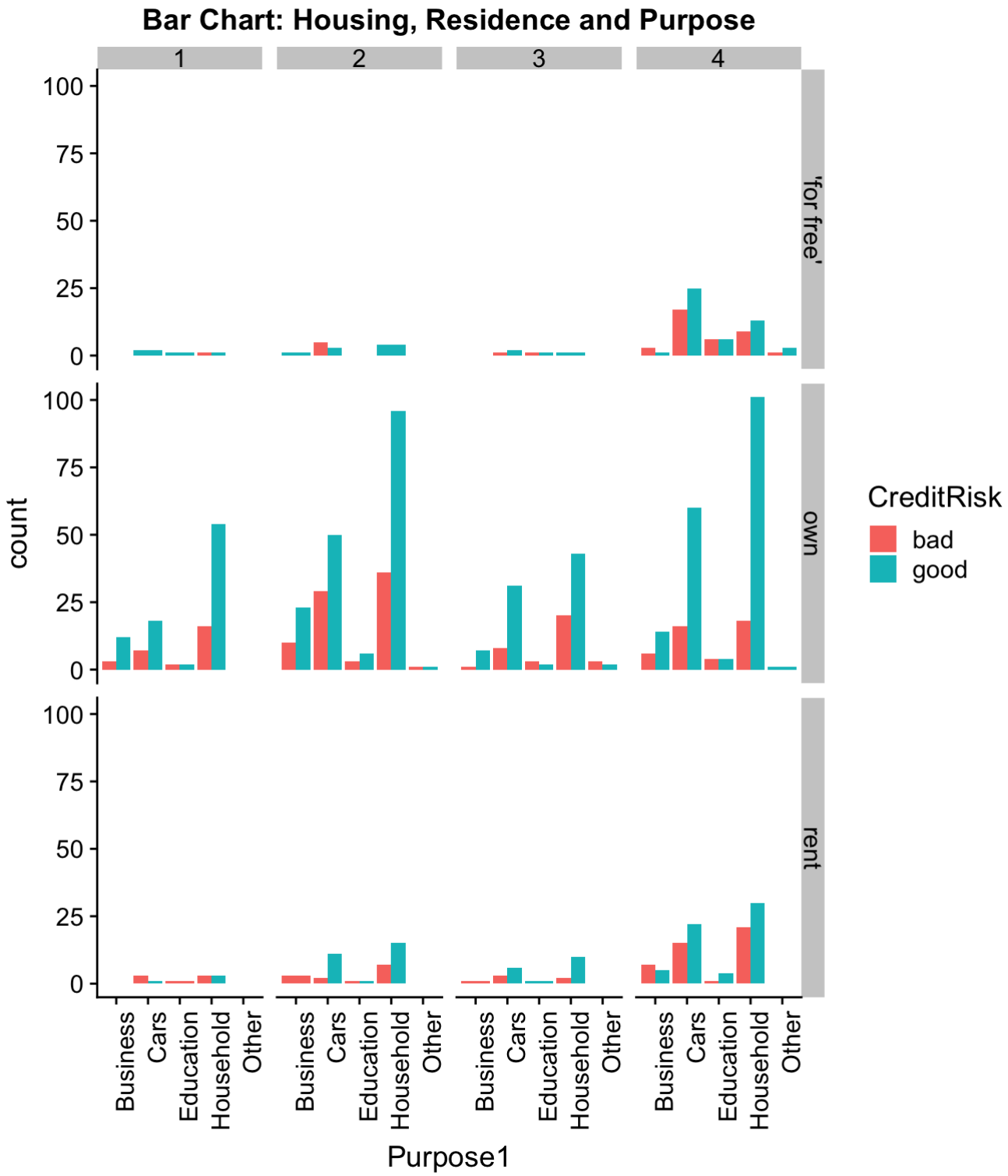
Job Type, Sex and Asset The following visual shows that most of the females had bad risk, and males had good risk. Unemplyed females however had good risk, whereas unemplyed males had bad risk. Most of the females had life insurance as an asset wheras males had real estate. The credit risk based on job type differed a lot between males and females. Particular aspects favoured females whereas others favoured males.



Age, Sex and Employment Years The following visualisation shows that the highest concentration of males varied from 20 to 55, wheras for females it varied between 20 and 45. There were also more males who were emplyed for more than 1 years. A lot of males had been employed for more than 7 years, as compared to females. Most of the females had less than 4 years of employment.

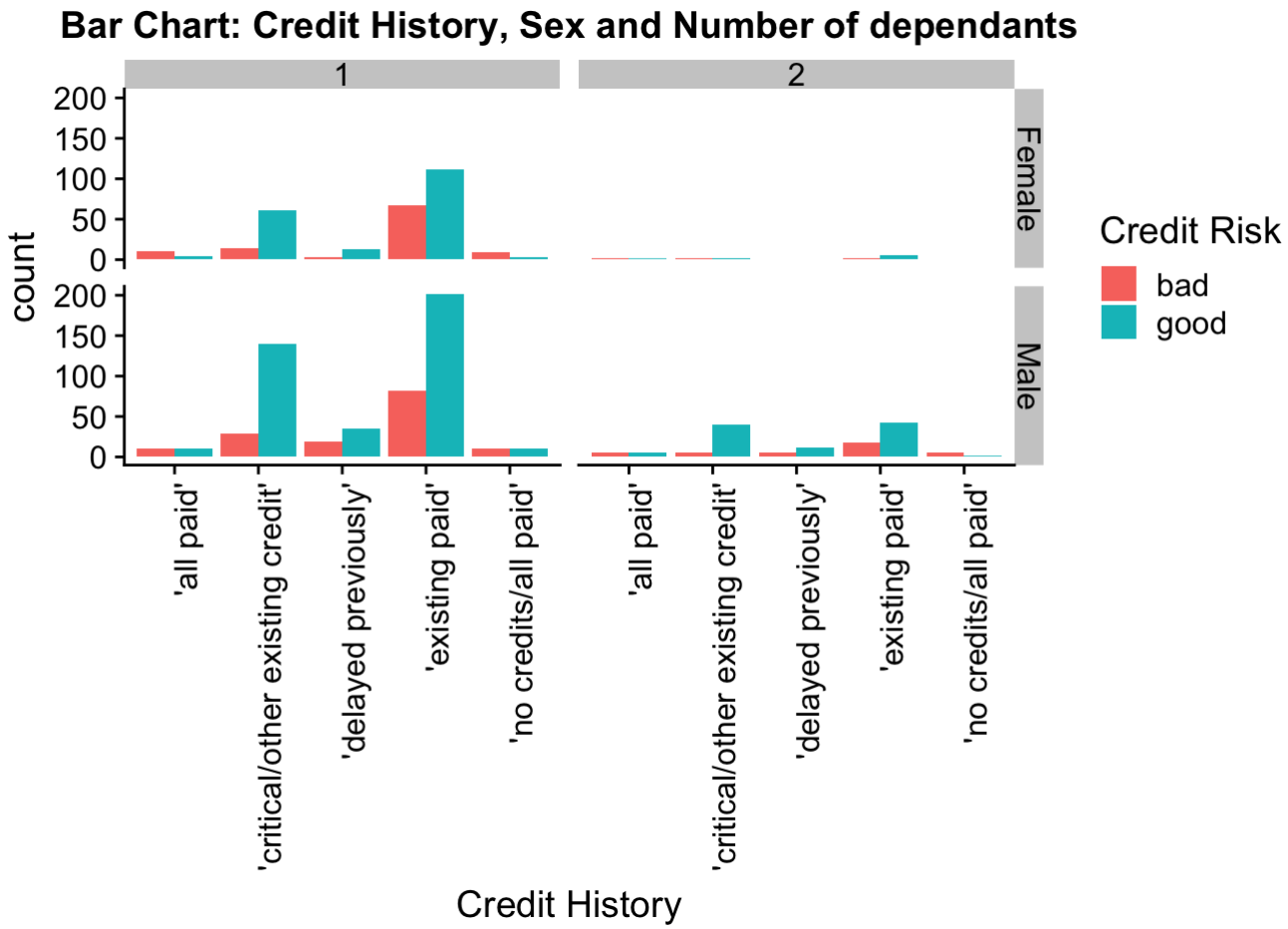


Housing, Residence and Purpose The following visualisation shows that most of the individuals had their own housing, out of which maximum had good credit. Lots of individuals who were staying for free for 4 years had bad credit risk. Credit for household expenses was the most sought after credit scheme for individuals with their own house, wheras people who lived for free had credit for cars. People on rent had similar numbers in credit for purpose of both cars and household.



Credit History, Sex and Number of dependants From the visualisation, we can see that most of the females had only 1 dependant, whereas males had slightly more in proportion. Males with 1 or 2 dependant mainly either had existing paid credit or critical credit.

```
ggplot(credit, aes(x = CreditHistory, fill = CreditRisk)) +
  geom_bar(position = 'dodge') + facet_grid( Status1~ Dependants ) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  labs(title = 'Bar Chart: Credit History, Sex and Number of dependants', x = "Credit
History", fill = "Credit Risk")
```



Summary

From the data exploration phase, we found that the numerical features were quite tidy, and hence no cleaning needed to be undertaken. In case of the categorical variables, we had to define some new features which binned their corresponding original features into lower cardinalities. We did not remove any of the original variables in the dataset, as we would like to play around with granularity at the model building phase.

From the above exploration and visualisation in the Data Exploration step, we found that Status1 (Gender), Purpose1 (Purpose of credit), Housing (Current Accomodation type), Age, Employment (In years), JobType and Asset (eg. property, insurance) can possiblly be useful predictors for the CreditRisk class.