Exploratory Data Analysis

2023-03-24

Load Required Libraries

```
library(dplyr)
library(insight)
library(knitr)
library(kableExtra)
library(ggplot2)
library(tidyverse)
library(corrplot)
library(patchwork)
```

Load Data & Inspect Variables

```
# Read the data
data <- read.csv("Credit.csv")
# Check the number of observations and number of variables
n <- nrow(data)
m <- ncol(data)
n
```

```
## [1] 1000
```

```
m
```

```
## [1] 21
```

```
# Check the data
kable(head(data[, 1:8]), format = "latex", align=rep("c", 8))
```

status	duration	credit_history	purpose	amount	savings	employment_duration	installment_rate
1	18	4	2	1049	1	2	4
1	9	4	0	2799	1	3	2
2	12	2	9	841	2	4	2
1	12	4	0	2122	1	3	3
1	12	4	0	2171	1	3	4
1	10	4	0	2241	1	2	1

```
kable(head(data[, 9:14]), format = "latex", align=rep("c", 6))
```

personal_status_sex	other_debtors	present_residence	property	age	other_installment_plans
2	1	4	2	21	3
3	1	2	1	36	3
2	1	4	1	23	3
3	1	2	1	39	3
3	1	4	2	38	1
3	1	3	1	48	3

```
kable(head(data[, 15:21]), format = "latex", align=rep("c", 7))
```

housing	number_credits	job	people_liable	telephone	foreign_worker	credit_risk
1	1	3	2	1	2	1
1	2	3	1	1	2	1
1	1	2	2	1	2	1
1	2	2	1	1	1	1
2	2	2	2	1	1	1
1	2	2	1	1	1	1

```
# Check invalid or missing values
anyNA(data)
```

```
## [1] FALSE
```

```
# Check the data type of each column
sapply(data, class)
```

```
##                  status                duration          credit_history
##               "integer"               "integer"               "integer"
##                 purpose                  amount                 savings
##               "integer"               "integer"               "integer"
##     employment_duration        installment_rate     personal_status_sex
##               "integer"               "integer"               "integer"
##           other_debtors       present_residence                property
##               "integer"               "integer"               "integer"
##                     age  other_installment_plans                 housing
##               "integer"               "integer"               "integer"
##          number_credits                     job           people_liable
##               "integer"               "integer"               "integer"
##               telephone          foreign_worker             credit_risk
##               "integer"               "integer"               "integer"
```

As we can see from the above outputs, there is no NaN values so the data is clean. And all of the columns are of type integer. Some of them are quantitative variable while some of them are qualitative variables. Here is a summary of the variables:

- status: status of the debtor's checking account with the bank (categorical)
- duration: credit duration in months (quantitative)
- credit_history: history of compliance with previous or concurrent credit contracts (categorical)
- purpose: purpose for which the credit is needed (categorical)
- amount: credit amount in DM (quantitative; result of monotonic transformation; actual data and type of transformation unknown)
- savings: debtor's savings (categorical)

- employment_duration: duration of debtor's employment with current employer (ordinal; discretized quantitative)
- installment_rate: credit installments as a percentage of debtor's disposable income (ordinal; discretized quantitative)
- personal_status_sex: combined information on sex and marital status (categorical)
- other_debtors: is there another debtor or a guarantor for the credit? (categorial)
- present_residence: length of time (in years) the debtor lives in the present residence (ordinal; discretized quantitative)
- property: the debtor's most valuable property (ordinal)
- age: age in years (quantitative)
- other_installment_plans: installment plans from providers other than the credit-giving bank (categorical)
- housing: type of housing the debtor lives in (categorical)
- number_credits: number of credits including the current one the debtor has (or had) at the bank (ordinal; discretized quantitative)
- job: quality of debtor's job (ordinal)
- people_liable: number of persons who financially depend on the debtor (binary; discretized quantitative)
- telephone: is there a telephone landline registered on the debtor's name? (binary)
- foreign_ worker: is the debtor a foreign worker? (binary)
- credit_risk: has the credit contract been complied with (good) or not (bad)? (binary)

We can see that the **quantitative variables** include duration, amount and age, while **qualitative variables** include status, credit_history, purpose, savings, employment_duration, installment_rate, personal_status_sex, other_debtors, present_residence, property, other_installment_plans, housing, number_credits, job, people_liable, telephone, foreign_worker and credit_risk.

Univariate Analysis & Visualization

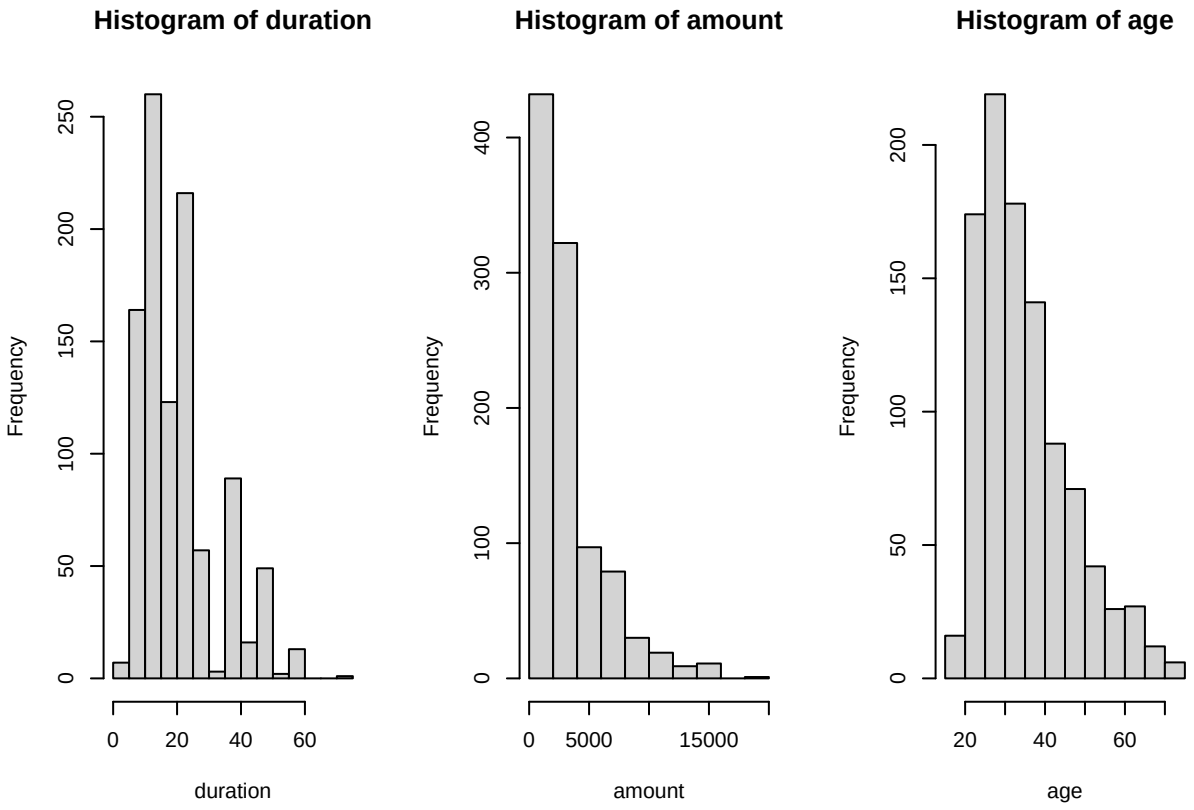
First we will perform univariate analysis on each of the variables and look at their distribution. Here is the summary statistics:

```
quant_vars <- c("duration", "amount", "age")
qual_vars <- c("status", "credit_history", "purpose", "savings", "employment_duration",
               "installment_rate", "personal_status_sex", "other_debtors", "present_residence",
               "property", "other_installment_plans", "housing", "number_credits", "job", "people_liabl
               "telephone", "foreign_worker")
summary(data[, quant_vars])
```

```
##     duration        amount           age
##  Min.   : 4.0   Min.   :  250   Min.   :19.00
##  1st Qu.:12.0   1st Qu.: 1366   1st Qu.:27.00
##  Median :18.0   Median : 2320   Median :33.00
##  Mean   :20.9   Mean   : 3271   Mean   :35.54
##  3rd Qu.:24.0   3rd Qu.: 3972   3rd Qu.:42.00
##  Max.   :72.0   Max.   :18424   Max.   :75.00
```

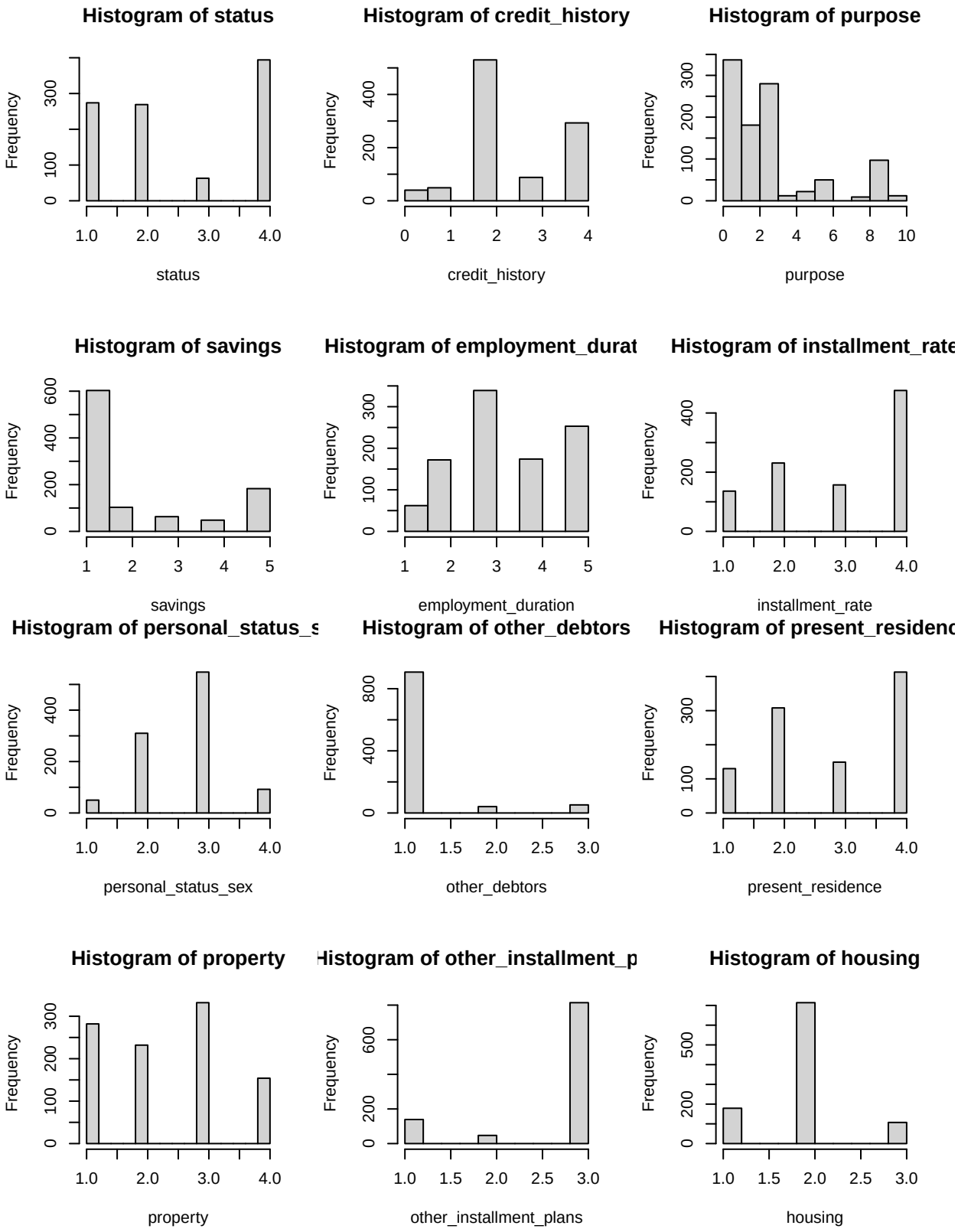
Next, let us check the histograms of the quantitative variables:

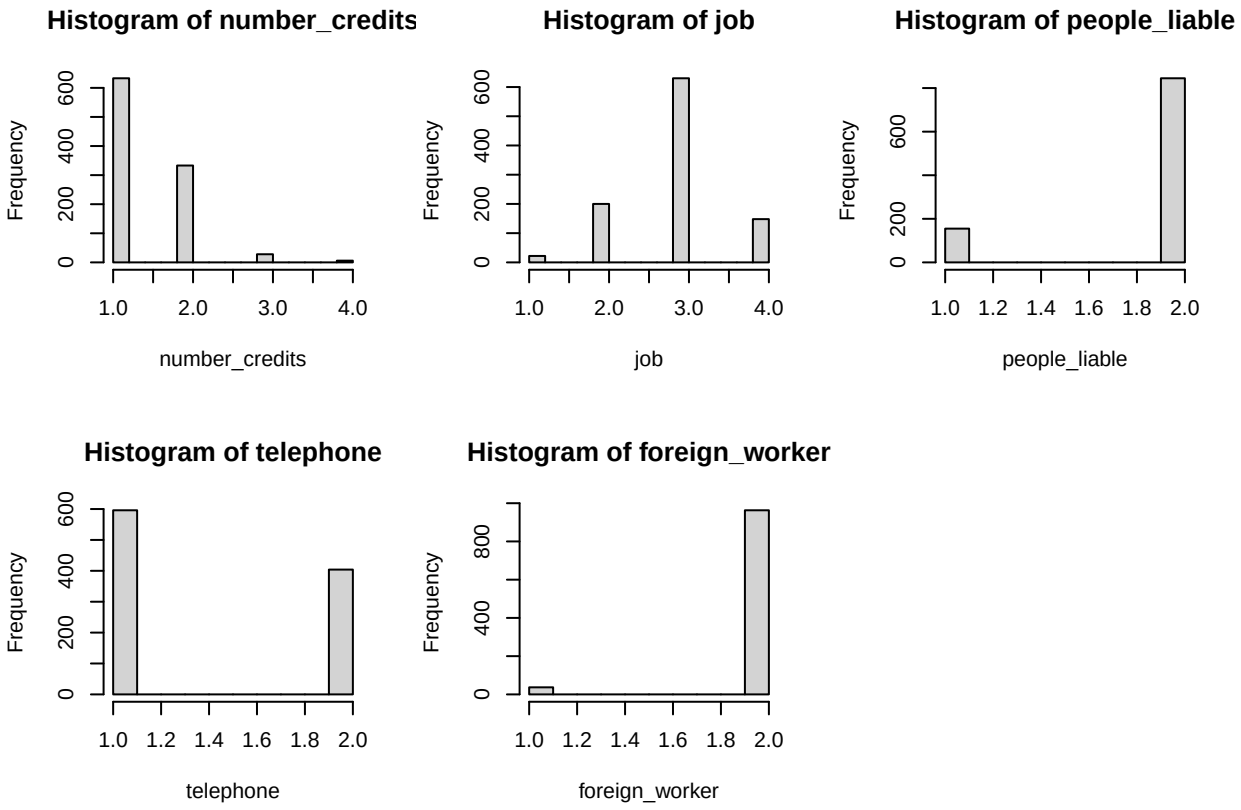
```
par(mfrow=c(1,3))
lapply(quant_vars, FUN=function(c)
  hist(data[, c], xlab=c, main=paste("Histogram of", c)))
```



Then, let us check the barplots of qualitative variables:

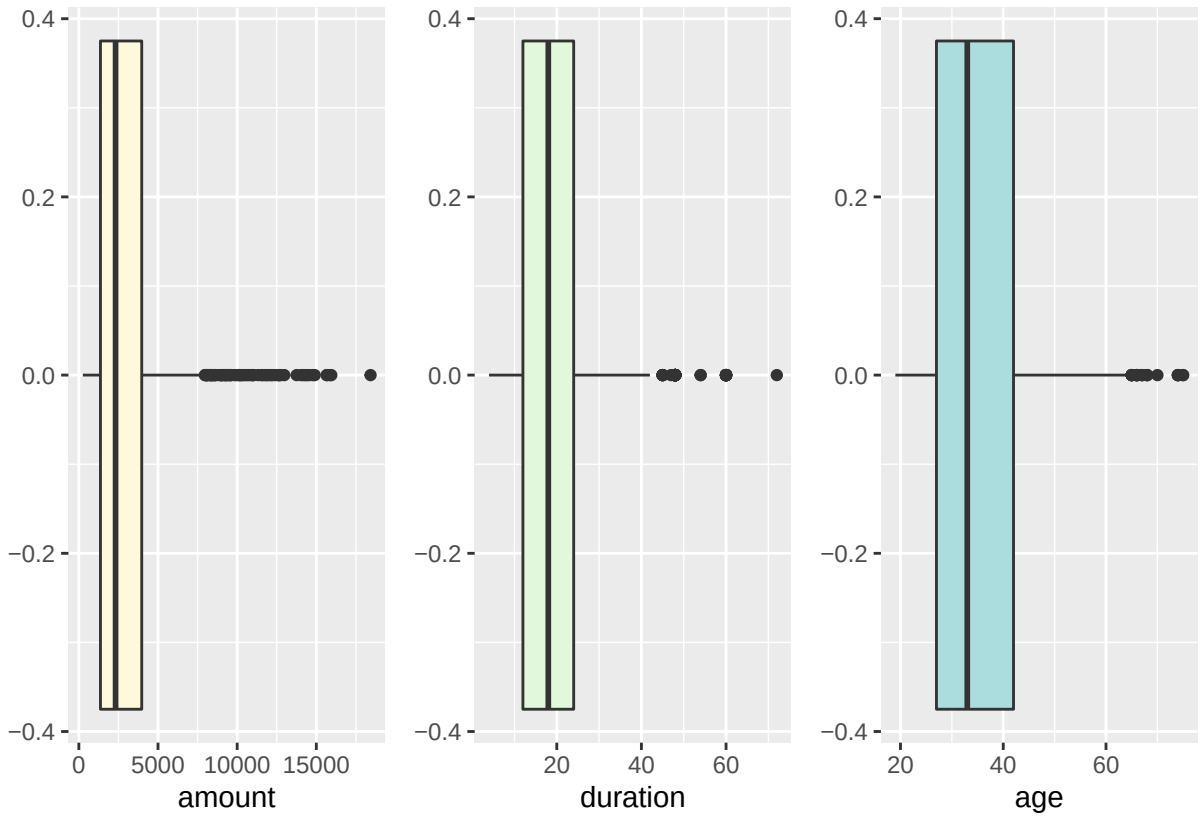
```
par(mfrow=c(2, 3))
lapply(qual_vars, FUN=function(c)hist(data[, c], xlab=c, main=paste("Histogram of", c)))
```


Histogram of status
Frequency
0
100
300
1.0
2.0
3.0
4.0
status
Histogram of credit_history
Frequency
0
200
400
0
1
2
3
4
credit_history
Histogram of purpose
Frequency
0
100
200
300
0
2
4
6
8
10
purpose
Histogram of savings
Frequency
0
200
400
600
1
2
3
4
5
savings
Histogram of employment_durat
Frequency
0
100
200
300
1
2
3
4
5
employment_duration
Histogram of installment_rate
Frequency
0
200
400
1.0
2.0
3.0
4.0
installment_rate
Histogram of personal_status_s
Frequency
0
200
400
1.0
2.0
3.0
4.0
personal_status_sex
Histogram of other_debtors
Frequency
0
400
800
1.0
1.5
2.0
2.5
3.0
other_debtors
Histogram of present_residenc
Frequency
0
100
300
1.0
2.0
3.0
4.0
present_residence
Histogram of property
Frequency
0
100
200
300
1.0
2.0
3.0
4.0
property
Histogram of other_installment_p
Frequency
0
200
600
1.0
1.5
2.0
2.5
3.0
other_installment_plans
Histogram of housing
Frequency
0
200
500
1.0
1.5
2.0
2.5
3.0
housing



After checking the histograms and barplots, we will check the boxplots of the quantitative variables. Here we will not check barplots for qualitative variables because it only makes sense to examine the median, first and third quartiles and maximum value for quantitative variables.

```
g1 <- ggplot(data, aes(x = amount)) + geom_boxplot(fill="#FEF8DD")
g2 <- ggplot(data, aes(x = duration)) + geom_boxplot(fill="#E1F8DC")
g3 <- ggplot(data, aes(x = age)) + geom_boxplot(fill="#ACDDDE")
g1 + g2 + g3
```



From the above box plots, we can see that there are a few outliers for the variable amount. If we look at the histogram of variable amount, we can see that it is a right skewed distribution with a long right tail, which results in these outliers.

For binary variables people_liable, telephone, foreign_worker and credit_risk, we can calculate and interpret the sample odds:

```
binary_var <- c("Statistics", "people_liable", "telephone", "foreign_worker", "credit_risk")
odds <- c("Sample Odds")
for (var in binary_var[2:5]) {
  if (var == "credit_risk") {y <- sum(data[, var] == 1)}
  else {y <- sum(data[, var] == 2)}
  n <- length(data[, var])
  odds <- append(odds, round(y / (n - y), 2))
}
kable(data.frame(t(odds)), col.names = binary_var, format = "latex") %>%
  kable_styling(position = "center", latex_options = "hold_position") %>% row_spec(0, bold = TRUE)
```

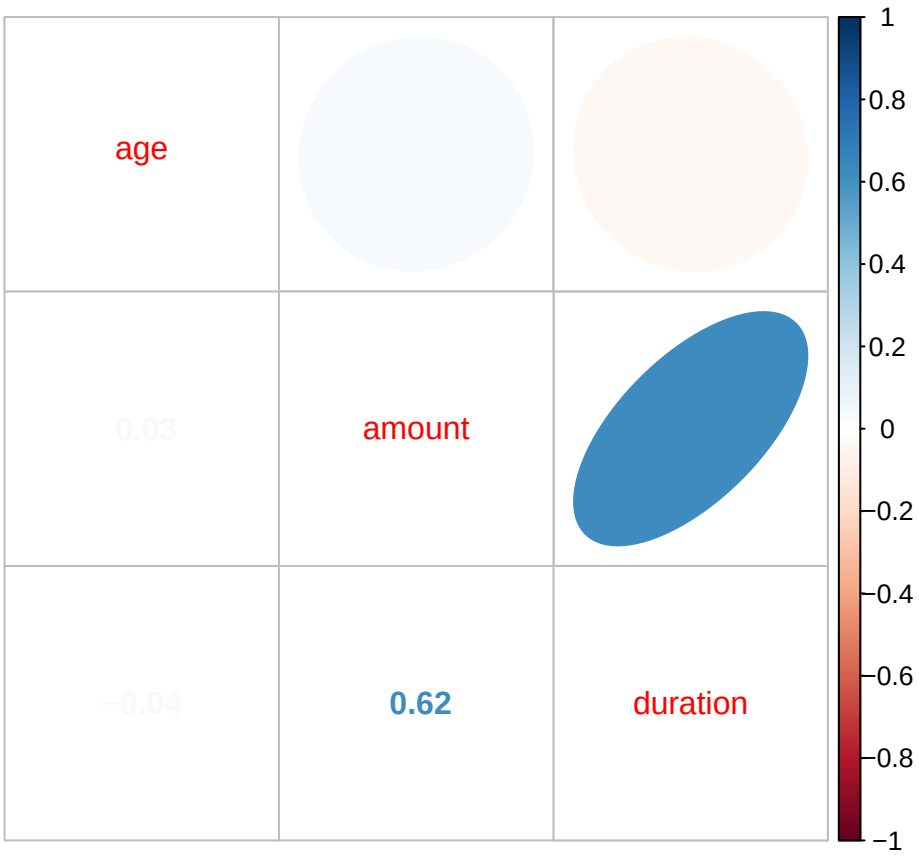
Statistics	people_liable	telephone	foreign_worker	credit_risk
Sample Odds	5.45	0.68	26.03	2.33

Based on our sample, the estimated probability of a person to have good credit is 2.33 times as likely as having a bad credit. Similarly, the estimated probability of a person to have a telephone landline registered on his/her name is 0.68 times as likely as not having such a telephone landline.

Multivariate Analysis

First, let us look at the correlation plots of the quantitative variables.

```
corrplot.mixed(cor(data[quant_vars]), lower='number', upper='ellipse', order='AOE')
```



From the above correlation plot, we can see a strong correlation between amount and duration.