Exploratory Data Analysis

2023-03-22

Load Required Libraries

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

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:

```
summary(data)
```

```
##      status         duration    credit_history     purpose
##  Min.   :1.000   Min.   : 4.0   Min.   :0.000   Min.   : 0.000
##  1st Qu.:1.000   1st Qu.:12.0   1st Qu.:2.000   1st Qu.: 1.000
##  Median :2.000   Median :18.0   Median :2.000   Median : 2.000
##  Mean   :2.577   Mean   :20.9   Mean   :2.545   Mean   : 2.828
##  3rd Qu.:4.000   3rd Qu.:24.0   3rd Qu.:4.000   3rd Qu.: 3.000
##  Max.   :4.000   Max.   :72.0   Max.   :4.000   Max.   :10.000
##      amount         savings      employment_duration installment_rate
##  Min.   :  250   Min.   :1.000   Min.   :1.000       Min.   :1.000
##  1st Qu.: 1366   1st Qu.:1.000   1st Qu.:3.000       1st Qu.:2.000
##  Median : 2320   Median :1.000   Median :3.000       Median :3.000
##  Mean   : 3271   Mean   :2.105   Mean   :3.384       Mean   :2.973
##  3rd Qu.: 3972   3rd Qu.:3.000   3rd Qu.:5.000       3rd Qu.:4.000
##  Max.   :18424   Max.   :5.000   Max.   :5.000       Max.   :4.000
##  personal_status_sex other_debtors   present_residence    property
##  Min.   :1.000       Min.   :1.000   Min.   :1.000     Min.   :1.000
##  1st Qu.:2.000       1st Qu.:1.000   1st Qu.:2.000     1st Qu.:1.000
##  Median :3.000       Median :1.000   Median :3.000     Median :2.000
```

```
##  Mean   :2.682       Mean   :1.145   Mean   :2.845     Mean   :2.358
##  3rd Qu.:3.000       3rd Qu.:1.000   3rd Qu.:4.000     3rd Qu.:3.000
##  Max.   :4.000       Max.   :3.000   Max.   :4.000     Max.   :4.000
##       age         other_installment_plans    housing      number_credits
##  Min.   :19.00   Min.   :1.000           Min.   :1.000   Min.   :1.000
##  1st Qu.:27.00   1st Qu.:3.000           1st Qu.:2.000   1st Qu.:1.000
##  Median :33.00   Median :3.000           Median :2.000   Median :1.000
##  Mean   :35.54   Mean   :2.675           Mean   :1.928   Mean   :1.407
##  3rd Qu.:42.00   3rd Qu.:3.000           3rd Qu.:2.000   3rd Qu.:2.000
##  Max.   :75.00   Max.   :3.000           Max.   :3.000   Max.   :4.000
##       job         people_liable     telephone     foreign_worker   credit_risk
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :0.0
##  1st Qu.:3.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:0.0
##  Median :3.000   Median :2.000   Median :1.000   Median :2.000   Median :1.0
##  Mean   :2.904   Mean   :1.845   Mean   :1.404   Mean   :1.963   Mean   :0.7
##  3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:1.0
##  Max.   :4.000   Max.   :2.000   Max.   :2.000   Max.   :2.000   Max.   :1.0
```

Next, let us check the histograms of the variables:

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