projek fiksss

Nabella Yunita Sari_164231019

2024-12-10

Import Library

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(readr)
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(VIM)
```

```
## Loading required package: colorspace

## Loading required package: grid

## VIM is ready to use.

## Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues
```

```
##
## Attaching package: 'VIM'

## The following object is masked from 'package:datasets':
##
##     sleep
```

```r
library(mice)
```

```
##
## Attaching package: 'mice'

## The following object is masked from 'package:stats':
##
##     filter

## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

```r
library(tidyr)
library(caret)
```

```
## Loading required package: lattice
```

```r
library(FactoMineR)
```

```
## Warning: package 'FactoMineR' was built under R version 4.4.2
```

```r
library(factoextra)
```

```
## Warning: package 'factoextra' was built under R version 4.4.2

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

Load Data

```r
data_loan <- read_csv("C:/Users/HP/Downloads/LoanData_Raw_v1.0.csv")
```

```
## Rows: 700 Columns: 9
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): default
## dbl (8): age, ed, employ, address, income, debtinc, creddebt, othdebt
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
head(data)
```

```
##
## 1 function (..., list = character(), package = NULL, lib.loc = NULL,
## 2     verbose = getOption("verbose"), envir = .GlobalEnv, overwrite = TRUE)
## 3 {
## 4     fileExt <- function(x) {
## 5         db <- grepl("\\\\.[^.]+\\\\.(gz|bz2|xz)$", x)
## 6         ans <- sub(".*\\\\.", "", x)
```

Cek Kualitas Data

Clean Data

```r
data_loan$default <- as.character(data_loan$default)
data_loan$default <- ifelse(data_loan$default %in% c("'0'", ":0", "0"), 0, 1)
data_loan$default <- as.factor(data_loan$default)
```

Banyak Baris dan Kolom

```r
nrow(data_loan)
```

```
## [1] 700
```

```r
ncol(data_loan)
```

```
## [1] 9
```

Tipe Data

```r
str(data_loan)
```

```
## spc_tbl_ [700 x 9] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ age     : num [1:700] 41 27 40 41 24 41 39 NA 24 36 ...
##  $ ed      : num [1:700] 3 1 1 NA 2 2 1 1 1 1 ...
##  $ employ  : num [1:700] 17 10 15 15 2 5 20 12 3 0 ...
##  $ address : num [1:700] 12 6 7 14 0 5 9 11 4 13 ...
##  $ income  : num [1:700] 176 31 NA 120 28 25 NA 38 19 25 ...
##  $ debtinc : num [1:700] 9.3 17.3 5.5 2.9 17.3 10.2 30.6 3.6 24.4 19.7 ...
##  $ creddebt: num [1:700] 11.359 1.362 0.856 2.659 1.787 ...
##  $ othdebt : num [1:700] 5.009 4.001 2.169 0.821 3.057 ...
##  $ default : Factor w/ 2 levels "0","1": 2 1 1 1 2 1 1 1 2 1 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   age = col_double(),
```

```
##   ..   ed = col_double(),
##   ..   employ = col_double(),
##   ..   address = col_double(),
##   ..   income = col_double(),
##   ..   debtinc = col_double(),
##   ..   creddebt = col_double(),
##   ..   othdebt = col_double(),
##   ..   default = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

Cek Jumlah Unique untuk Setiap Kolom

```r
jumlah_unique <- sapply(data_loan, function(x) length(unique(x)))
jumlah_unique
```

```
##      age       ed   employ  address   income  debtinc creddebt  othdebt 
##       39        6       32       31      114      231      695      699 
##  default 
##        2
```

Cek Duplikasi Data

```r
duplicates <- data_loan %>%
  filter(duplicated(.))
print(paste("Jumlah baris duplikat:", nrow(duplicates)))
```

```
## [1] "Jumlah baris duplikat: 0"
```

Ringkasan Data

```r
summary(data_loan)
```

```
##       age              ed            employ          address      
##  Min.   : 20.0   Min.   :1.000   Min.   : 0.000   Min.   : 0.000  
##  1st Qu.: 28.0   1st Qu.:1.000   1st Qu.: 3.000   1st Qu.: 3.000  
##  Median : 34.0   Median :1.000   Median : 7.000   Median : 7.000  
##  Mean   : 34.9   Mean   :1.718   Mean   : 8.389   Mean   : 8.269  
##  3rd Qu.: 40.0   3rd Qu.:2.000   3rd Qu.:12.000   3rd Qu.:12.000  
##  Max.   :136.0   Max.   :5.000   Max.   :31.000   Max.   :34.000  
##  NA's   :19      NA's   :20                                       
##      income          debtinc          creddebt           othdebt           default
##  Min.   : 14.00   Min.   : 0.40   Min.   : 0.0117   Min.   : 0.04558   0:517  
##  1st Qu.: 24.00   1st Qu.: 5.00   1st Qu.: 0.3691   1st Qu.: 1.04418   1:183  
##  Median : 34.00   Median : 8.60   Median : 0.8549   Median : 1.98757          
##  Mean   : 45.74   Mean   :10.26   Mean   : 1.5536   Mean   : 3.05821          
##  3rd Qu.: 54.50   3rd Qu.:14.12   3rd Qu.: 1.9020   3rd Qu.: 3.92306          
##  Max.   :446.00   Max.   :41.30   Max.   :20.5613   Max.   :27.03360          
##  NA's   :37
```

Outliers

Cek Jumlah Outlier

```r
count_outliers <- function(column) {
  Q1 <- quantile(column, 0.25, na.rm = TRUE)
  Q3 <- quantile(column, 0.75, na.rm = TRUE)
  IQR <- Q3 - Q1
  lower_bound <- Q1 - 1.5 * IQR
  upper_bound <- Q3 + 1.5 * IQR
  sum(column < lower_bound | column > upper_bound, na.rm = TRUE)
}

outliers_per_column <- sapply(data_loan, function(col) {
  if (is.numeric(col)) {
    count_outliers(col)
  } else {
    NA
  }
})

outliers_per_column
```

```
##      age       ed   employ  address   income  debtinc creddebt  othdebt
##        1       41       10       14       43       14       55       48
##  default
##       NA
```

```r
percent_outliers <- function(column) {
  Q1 <- quantile(column, 0.25, na.rm = TRUE)
  Q3 <- quantile(column, 0.75, na.rm = TRUE)
  IQR <- Q3 - Q1
  lower_bound <- Q1 - 1.5 * IQR
  upper_bound <- Q3 + 1.5 * IQR
  outlier_count <- sum(column < lower_bound | column > upper_bound, na.rm = TRUE)
  total_count <- sum(!is.na(column)) # Count of non-missing values
  (outlier_count / total_count) * 100 # Return percentage
}

percent_outliers_per_column <- sapply(data_loan, function(col) {
  if (is.numeric(col)) {
    percent_outliers(col)
  } else {
    NA # Skip non-numeric columns
  }
})

percent_outliers_per_column
```

```
##       age        ed    employ   address    income   debtinc  creddebt   othdebt
## 0.1468429 6.0294118 1.4285714 2.0000000 6.4856712 2.0000000 7.8571429 6.8571429
##   default
##        NA
```

Box plot masing-masing Variabel

Handling Outlier

```
# Function to handle outliers by replacing them with lower or upper bound
handle_outliers <- function(column) {
  Q1 <- quantile(column, 0.25, na.rm = TRUE)
  Q3 <- quantile(column, 0.75, na.rm = TRUE)
  IQR <- Q3 - Q1
  lower_bound <- Q1 - 1.5 * IQR
  upper_bound <- Q3 + 1.5 * IQR

  # Replace outliers with the lower or upper bound
  column[column < lower_bound] <- lower_bound
  column[column > upper_bound] <- upper_bound

  return(column)
}

numeric_columns <- names(data_loan)[sapply(data_loan, is.numeric)]

for (col in numeric_columns) {
  data_loan[[col]] <- handle_outliers(data_loan[[col]])
}

# View the data after outlier handling
head(data_loan)
```

```
## # A tibble: 6 x 9
##     age    ed employ address income debtinc creddebt othdebt default
##   <dbl> <dbl>  <dbl>   <dbl>  <dbl>   <dbl>    <dbl>   <dbl> <fct>
## 1    41     3     17      12   100.     9.3    4.20    5.01  1
## 2    27     1     10       6    31     17.3    1.36    4.00  0
## 3    40     1     15       7    NA      5.5    0.856   2.17  0
## 4    41    NA     15      14   100.     2.9    2.66    0.821 0
## 5    24     2      2       0    28     17.3    1.79    3.06  1
## 6    41     2      5       5    25     10.2    0.393   2.16  0
```

Cek Jumlah Outlier Setelah Handling

```
count_outliers <- function(column) {
  Q1 <- quantile(column, 0.25, na.rm = TRUE)
  Q3 <- quantile(column, 0.75, na.rm = TRUE)
  IQR <- Q3 - Q1
  lower_bound <- Q1 - 1.5 * IQR
  upper_bound <- Q3 + 1.5 * IQR
  sum(column < lower_bound | column > upper_bound, na.rm = TRUE)
}

outliers_per_column <- sapply(data_loan, function(col) {
```

```
  if (is.numeric(col)) {
    count_outliers(col)
  } else {
    NA
  }
})

outliers_per_column
```

```
##      age       ed   employ  address   income  debtinc creddebt  othdebt
##        0        0        0        0        0        0        0        0
##  default
##       NA
```

Cek Jumlah Missing Value untuk Setiap Kolom

```
jumlah_misval <- sapply(data_loan, function(x) sum(is.na(x)))
jumlah_misval
```

```
##      age       ed   employ  address   income  debtinc creddebt  othdebt
##       19       20        0        0       37        0        0        0
##  default
##        0
```

Persentase Missing Values Untuk Tiap Kolom

```
missing_values <- sapply(data_loan, function(x) sum(is.na(x)) / length(x) * 100)
missing_values
```

```
##      age       ed   employ  address   income  debtinc creddebt  othdebt
## 2.714286 2.857143 0.000000 0.000000 5.285714 0.000000 0.000000 0.000000
##  default
## 0.000000
```

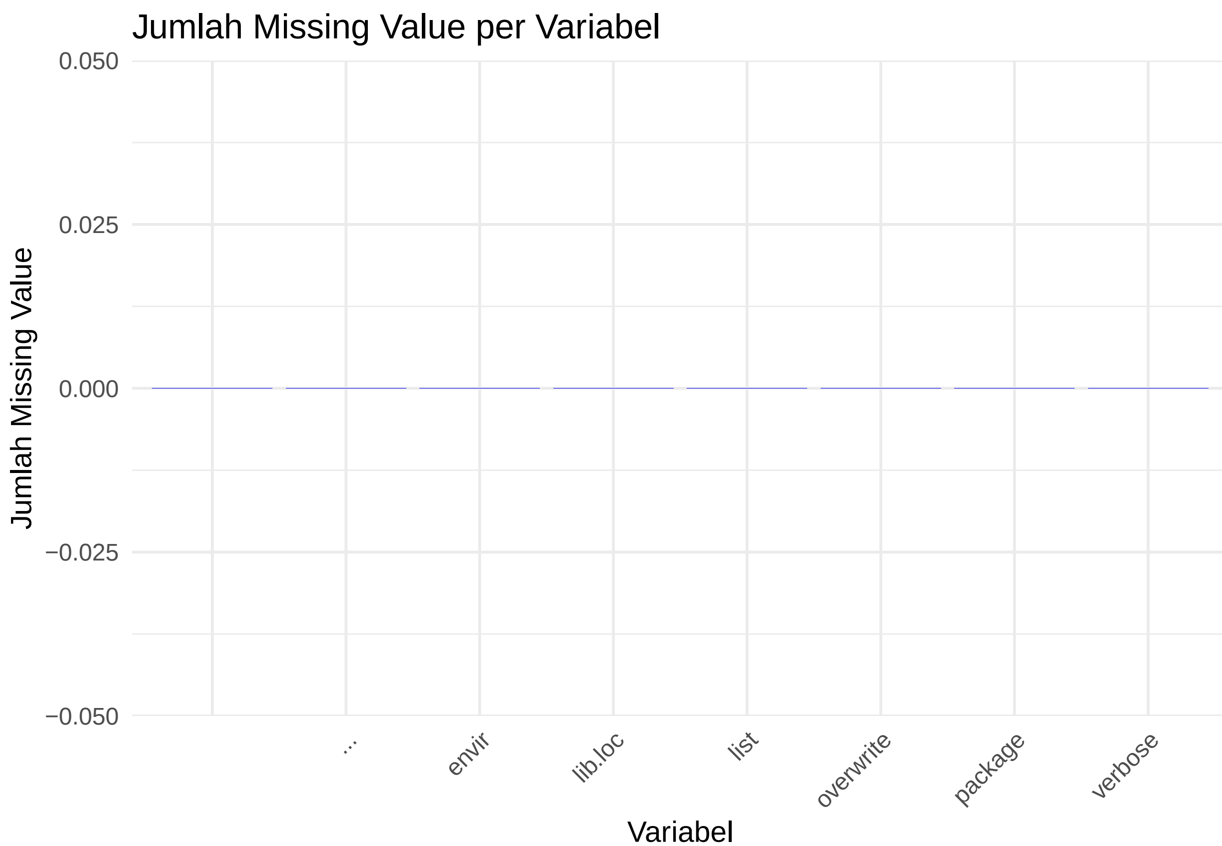
Bar Chart Missing Value

```
# Hitung jumlah missing value untuk setiap variabel
missing_data <- sapply(data, function(x) sum(is.na(x)))
```

```
## Warning in is.na(x): is.na() applied to non-(list or vector) of type 'symbol'

## Warning in is.na(x): is.na() applied to non-(list or vector) of type 'language'
## Warning in is.na(x): is.na() applied to non-(list or vector) of type 'language'

## Warning in is.na(x): is.na() applied to non-(list or vector) of type 'symbol'
```

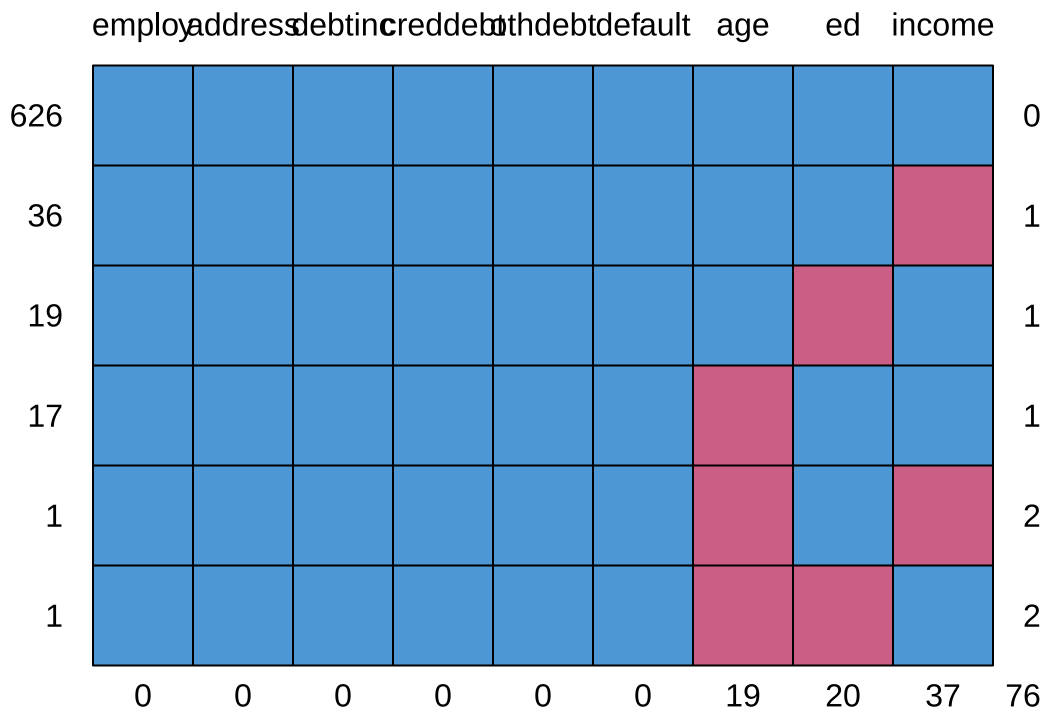
```
## Warning in is.na(x): is.na() applied to non-(list or vector) of type 'language'
```

```r
missing_data <- data.frame(Variable = names(missing_data), MissingValues = missing_data)

# Membuat bar chart
ggplot(missing_data, aes(x = reorder(Variable, MissingValues), y = MissingValues)) +
  geom_bar(stat = "identity", fill = "blue") +  # Warna biru
  labs(title = "Jumlah Missing Value per Variabel", x = "Variabel", y = "Jumlah Missing Value") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Cek Missing Values Berdasarkan Visualisasi dengan Mice

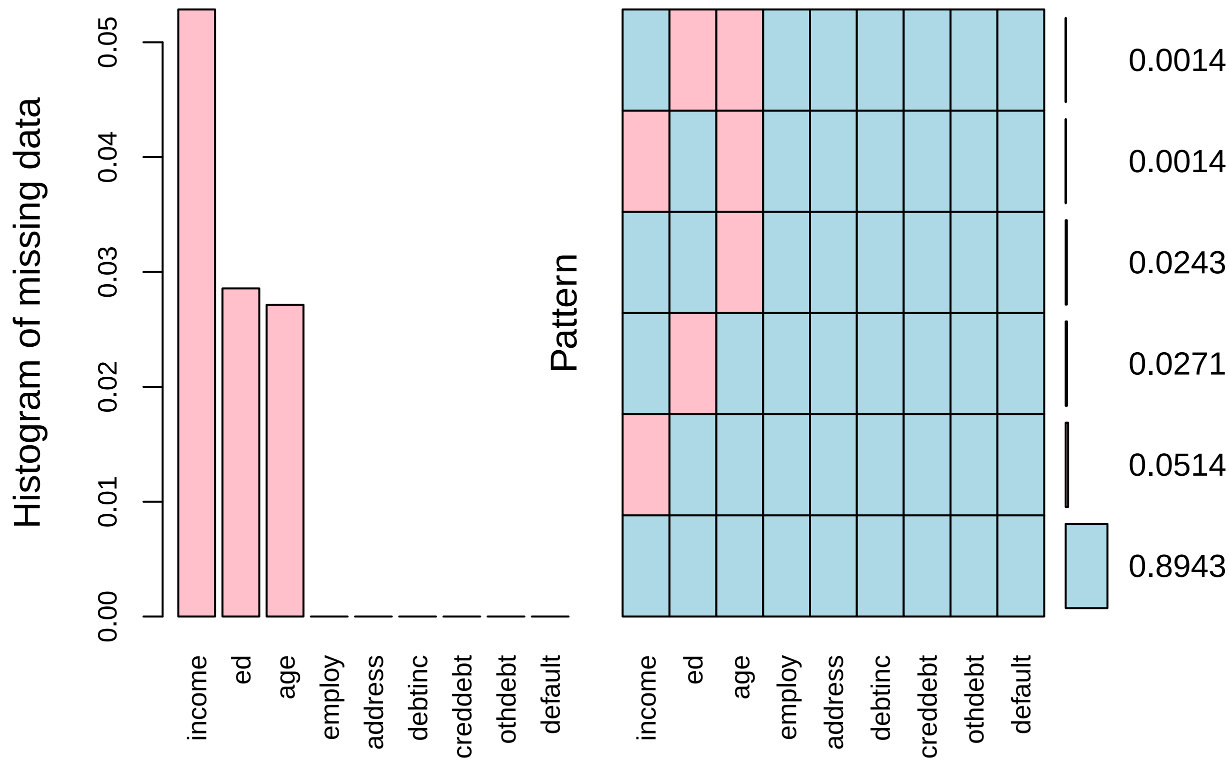
```r
library(mice)
md.pattern(data_loan)
```




```
##     employ address debtinc creddebt othdebt default age ed income
## 626      1       1       1        1       1       1   1  1      1  0
## 36       1       1       1        1       1       1   1  1      0  1
## 19       1       1       1        1       1       1   1  0      1  1
## 17       1       1       1        1       1       1   0  1      1  1
## 1        1       1       1        1       1       1   0  1      0  2
## 1        1       1       1        1       1       1   0  0      1  2
##          0       0       0        0       0       0  19 20     37 76
```

Cek Missing Values Berdasarkan Visualisasi dengan VIM

```r
library(VIM)
aggr_plot <- aggr(data_loan, col=c('lightblue','pink') , numbers=TRUE,
 sortVars=TRUE, labels=names(data_loan), cex.axis=.8,
 gap=1, ylab=c("Histogram of missing data","Pattern"))
```




```
##
##  Variables sorted by number of missings:
##  Variable      Count
##    income 0.05285714
##        ed 0.02857143
##       age 0.02714286
##    employ 0.00000000
##   address 0.00000000
##   debtinc 0.00000000
##  creddebt 0.00000000
##   othdebt 0.00000000
##   default 0.00000000
```

Imputasi Data Missing

```
# Imputasi dengan Metoden PMM
imputed_data1 <- mice(data_loan, m=5, maxit=50, method='pmm', seed=123)
```

```
##
##  iter imp variable
##   1   1  age  ed  income
##   1   2  age  ed  income
##   1   3  age  ed  income
```

```
##   1   4  age  ed  income
##   1   5  age  ed  income
##   2   1  age  ed  income
##   2   2  age  ed  income
##   2   3  age  ed  income
##   2   4  age  ed  income
##   2   5  age  ed  income
##   3   1  age  ed  income
##   3   2  age  ed  income
##   3   3  age  ed  income
##   3   4  age  ed  income
##   3   5  age  ed  income
##   4   1  age  ed  income
##   4   2  age  ed  income
##   4   3  age  ed  income
##   4   4  age  ed  income
##   4   5  age  ed  income
##   5   1  age  ed  income
##   5   2  age  ed  income
##   5   3  age  ed  income
##   5   4  age  ed  income
##   5   5  age  ed  income
##   6   1  age  ed  income
##   6   2  age  ed  income
##   6   3  age  ed  income
##   6   4  age  ed  income
##   6   5  age  ed  income
##   7   1  age  ed  income
##   7   2  age  ed  income
##   7   3  age  ed  income
##   7   4  age  ed  income
##   7   5  age  ed  income
##   8   1  age  ed  income
##   8   2  age  ed  income
##   8   3  age  ed  income
##   8   4  age  ed  income
##   8   5  age  ed  income
##   9   1  age  ed  income
##   9   2  age  ed  income
##   9   3  age  ed  income
##   9   4  age  ed  income
##   9   5  age  ed  income
##   10   1  age  ed  income
##   10   2  age  ed  income
##   10   3  age  ed  income
##   10   4  age  ed  income
##   10   5  age  ed  income
##   11   1  age  ed  income
##   11   2  age  ed  income
##   11   3  age  ed  income
##   11   4  age  ed  income
##   11   5  age  ed  income
##   12   1  age  ed  income
##   12   2  age  ed  income
```

```
##   12   3  age  ed  income
##   12   4  age  ed  income
##   12   5  age  ed  income
##   13   1  age  ed  income
##   13   2  age  ed  income
##   13   3  age  ed  income
##   13   4  age  ed  income
##   13   5  age  ed  income
##   14   1  age  ed  income
##   14   2  age  ed  income
##   14   3  age  ed  income
##   14   4  age  ed  income
##   14   5  age  ed  income
##   15   1  age  ed  income
##   15   2  age  ed  income
##   15   3  age  ed  income
##   15   4  age  ed  income
##   15   5  age  ed  income
##   16   1  age  ed  income
##   16   2  age  ed  income
##   16   3  age  ed  income
##   16   4  age  ed  income
##   16   5  age  ed  income
##   17   1  age  ed  income
##   17   2  age  ed  income
##   17   3  age  ed  income
##   17   4  age  ed  income
##   17   5  age  ed  income
##   18   1  age  ed  income
##   18   2  age  ed  income
##   18   3  age  ed  income
##   18   4  age  ed  income
##   18   5  age  ed  income
##   19   1  age  ed  income
##   19   2  age  ed  income
##   19   3  age  ed  income
##   19   4  age  ed  income
##   19   5  age  ed  income
##   20   1  age  ed  income
##   20   2  age  ed  income
##   20   3  age  ed  income
##   20   4  age  ed  income
##   20   5  age  ed  income
##   21   1  age  ed  income
##   21   2  age  ed  income
##   21   3  age  ed  income
##   21   4  age  ed  income
##   21   5  age  ed  income
##   22   1  age  ed  income
##   22   2  age  ed  income
##   22   3  age  ed  income
##   22   4  age  ed  income
##   22   5  age  ed  income
##   23   1  age  ed  income
```

```
## 23 2 age ed income
## 23 3 age ed income
## 23 4 age ed income
## 23 5 age ed income
## 24 1 age ed income
## 24 2 age ed income
## 24 3 age ed income
## 24 4 age ed income
## 24 5 age ed income
## 25 1 age ed income
## 25 2 age ed income
## 25 3 age ed income
## 25 4 age ed income
## 25 5 age ed income
## 26 1 age ed income
## 26 2 age ed income
## 26 3 age ed income
## 26 4 age ed income
## 26 5 age ed income
## 27 1 age ed income
## 27 2 age ed income
## 27 3 age ed income
## 27 4 age ed income
## 27 5 age ed income
## 28 1 age ed income
## 28 2 age ed income
## 28 3 age ed income
## 28 4 age ed income
## 28 5 age ed income
## 29 1 age ed income
## 29 2 age ed income
## 29 3 age ed income
## 29 4 age ed income
## 29 5 age ed income
## 30 1 age ed income
## 30 2 age ed income
## 30 3 age ed income
## 30 4 age ed income
## 30 5 age ed income
## 31 1 age ed income
## 31 2 age ed income
## 31 3 age ed income
## 31 4 age ed income
## 31 5 age ed income
## 32 1 age ed income
## 32 2 age ed income
## 32 3 age ed income
## 32 4 age ed income
## 32 5 age ed income
## 33 1 age ed income
## 33 2 age ed income
## 33 3 age ed income
## 33 4 age ed income
## 33 5 age ed income
```

```
## 34 1 age ed income
## 34 2 age ed income
## 34 3 age ed income
## 34 4 age ed income
## 34 5 age ed income
## 35 1 age ed income
## 35 2 age ed income
## 35 3 age ed income
## 35 4 age ed income
## 35 5 age ed income
## 36 1 age ed income
## 36 2 age ed income
## 36 3 age ed income
## 36 4 age ed income
## 36 5 age ed income
## 37 1 age ed income
## 37 2 age ed income
## 37 3 age ed income
## 37 4 age ed income
## 37 5 age ed income
## 38 1 age ed income
## 38 2 age ed income
## 38 3 age ed income
## 38 4 age ed income
## 38 5 age ed income
## 39 1 age ed income
## 39 2 age ed income
## 39 3 age ed income
## 39 4 age ed income
## 39 5 age ed income
## 40 1 age ed income
## 40 2 age ed income
## 40 3 age ed income
## 40 4 age ed income
## 40 5 age ed income
## 41 1 age ed income
## 41 2 age ed income
## 41 3 age ed income
## 41 4 age ed income
## 41 5 age ed income
## 42 1 age ed income
## 42 2 age ed income
## 42 3 age ed income
## 42 4 age ed income
## 42 5 age ed income
## 43 1 age ed income
## 43 2 age ed income
## 43 3 age ed income
## 43 4 age ed income
## 43 5 age ed income
## 44 1 age ed income
## 44 2 age ed income
## 44 3 age ed income
## 44 4 age ed income
```

```
##   44   5  age  ed  income
##   45   1  age  ed  income
##   45   2  age  ed  income
##   45   3  age  ed  income
##   45   4  age  ed  income
##   45   5  age  ed  income
##   46   1  age  ed  income
##   46   2  age  ed  income
##   46   3  age  ed  income
##   46   4  age  ed  income
##   46   5  age  ed  income
##   47   1  age  ed  income
##   47   2  age  ed  income
##   47   3  age  ed  income
##   47   4  age  ed  income
##   47   5  age  ed  income
##   48   1  age  ed  income
##   48   2  age  ed  income
##   48   3  age  ed  income
##   48   4  age  ed  income
##   48   5  age  ed  income
##   49   1  age  ed  income
##   49   2  age  ed  income
##   49   3  age  ed  income
##   49   4  age  ed  income
##   49   5  age  ed  income
##   50   1  age  ed  income
##   50   2  age  ed  income
##   50   3  age  ed  income
##   50   4  age  ed  income
##   50   5  age  ed  income
```

Mengekstrak dataset yang sudah diimputasi

```r
completed_data1 <- complete(imputed_data1)
head(completed_data1)
```

```
##   age ed employ address income debtinc creddebt  othdebt default
## 1  41  3     17      12 100.25     9.3 4.201299 5.008608       1
## 2  27  1     10       6  31.00    17.3 1.362202 4.000798       0
## 3  40  1     15       7  44.00     5.5 0.856075 2.168925       0
## 4  41  2     15      14 100.25     2.9 2.658720 0.821280       0
## 5  24  2      2       0  28.00    17.3 1.787436 3.056564       1
## 6  41  2      5       5  25.00    10.2 0.392700 2.157300       0
```

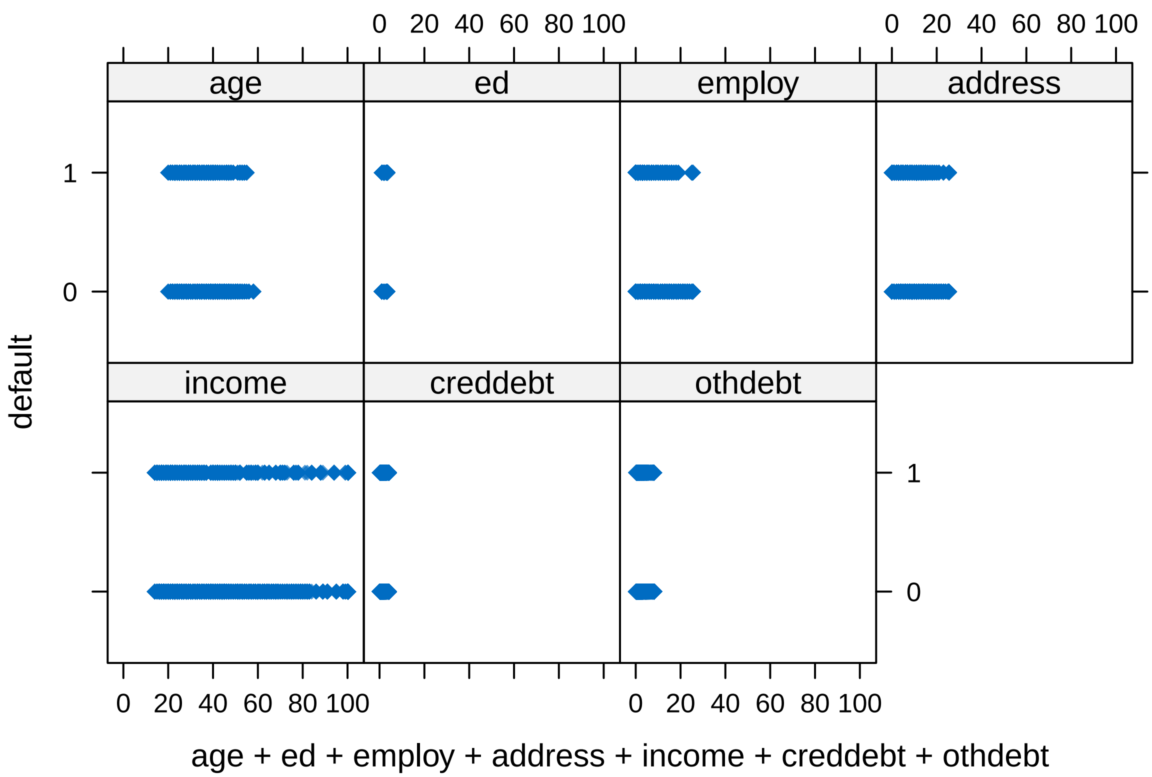
```r
summary(completed_data1)
```

```
##       age              ed            employ          address
##  Min.   :20.00   Min.   :1.000   Min.   : 0.000   Min.   : 0.000
##  1st Qu.:29.00   1st Qu.:1.000   1st Qu.: 3.000   1st Qu.: 3.000
##  Median :34.00   Median :1.000   Median : 7.000   Median : 7.000
```

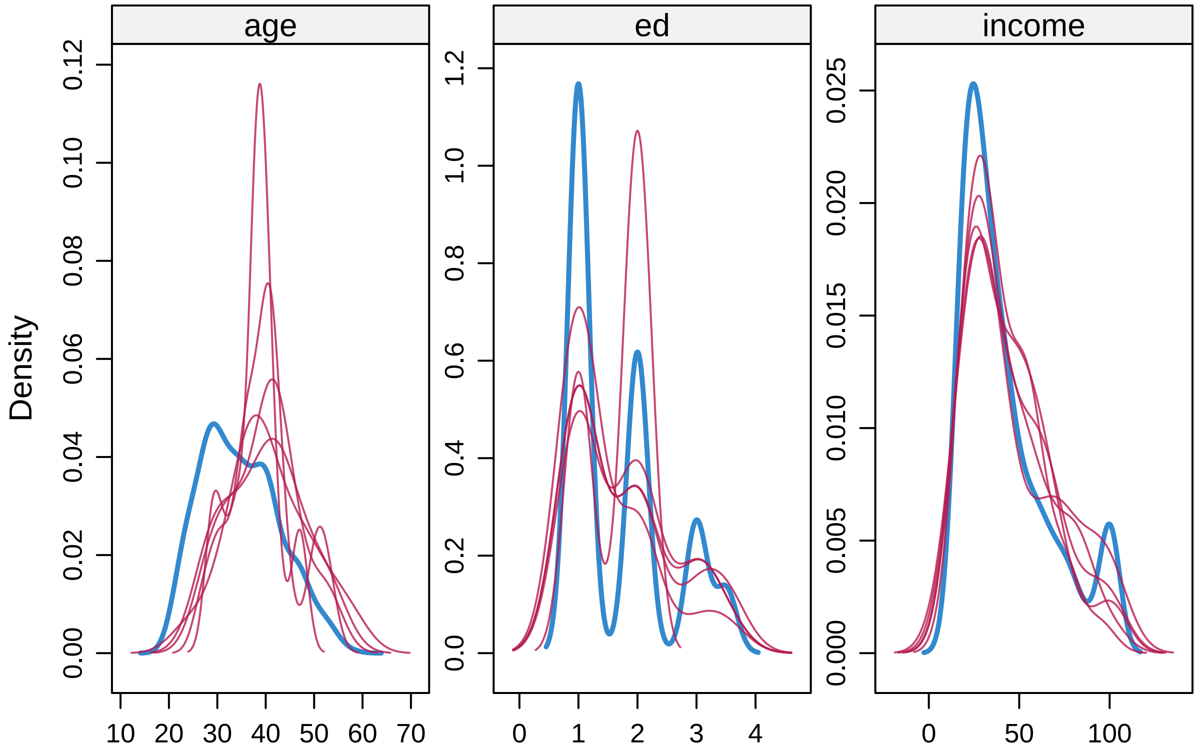
```
##  Mean   :34.94   Mean   :1.674   Mean   : 8.339   Mean   : 8.224  
##  3rd Qu.:41.00   3rd Qu.:2.000   3rd Qu.:12.000   3rd Qu.:12.000  
##  Max.   :58.00   Max.   :3.500   Max.   :25.500   Max.   :25.500  
##      income          debtinc          creddebt          othdebt         default
##  Min.   : 14.0   Min.   : 0.40   Min.   :0.0117   Min.   :0.04558   0:517  
##  1st Qu.: 24.0   1st Qu.: 5.00   1st Qu.:0.3691   1st Qu.:1.04418   1:183  
##  Median : 34.0   Median : 8.60   Median :0.8549   Median :1.98757          
##  Mean   : 42.4   Mean   :10.17   Mean   :1.3236   Mean   :2.76889          
##  3rd Qu.: 55.0   3rd Qu.:14.12   3rd Qu.:1.9020   3rd Qu.:3.92306          
##  Max.   :100.2   Max.   :27.81   Max.   :4.2013   Max.   :8.24139          
```

Visualisasi Imputasi Pertama dengan Metode pmm dengan maxit = 50

```
xyplot(imputed_data1,default ~ age+ed+employ+address+income+creddebt+othdebt,pch=18,cex=1)
```




```
densityplot(imputed_data1)
```



Cek Missing Value Setelah Imputasi

```
jumlah_misval_imputed <- sapply(completed_data1, function(x) sum(is.na(x)))
jumlah_misval_imputed
```

```
##      age       ed   employ  address   income  debtinc creddebt  othdebt
##        0        0        0        0        0        0        0        0
##  default
##        0
```

Cek Outliers Setelah Imputasi

```
count_outliers <- function(column) {
  Q1 <- quantile(column, 0.25, na.rm = TRUE)
  Q3 <- quantile(column, 0.75, na.rm = TRUE)
  IQR <- Q3 - Q1
  lower_bound <- Q1 - 1.5 * IQR
  upper_bound <- Q3 + 1.5 * IQR
  sum(column < lower_bound | column > upper_bound, na.rm = TRUE)
}

outliers_per_column <- sapply(completed_data1, function(col) {
  if (is.numeric(col)) {
    count_outliers(col)
```

```
  } else {
    NA
  }
})
outliers_per_column
```

```
##      age       ed   employ  address   income  debtinc creddebt  othdebt 
##        0        0        0        0        0        0        0        0 
##  default 
##       NA
```

```
cor_matrix <- cor(completed_data1[, c("age", "debtinc", "creddebt", "income", "othdebt", "employ", "addi
print(cor_matrix)
```

```
##                  age      debtinc  creddebt      income   othdebt      employ
## age      1.000000000  0.001244136 0.3183638  0.55901080 0.3579444  0.53490513
## debtinc  0.001244136  1.000000000 0.5865300 -0.01507331 0.6573873 -0.03817559
## creddebt 0.318363802  0.586529977 1.0000000  0.54513341 0.6543419  0.40014793
## income   0.559010803 -0.015073311 0.5451334  1.00000000 0.6009603  0.72330809
## othdebt  0.357944396  0.657387302 0.6543419  0.60096027 1.0000000  0.42513013
## employ   0.534905133 -0.038175592 0.4001479  0.72330809 0.4251301  1.00000000
## address  0.583554170  0.016412042 0.2309154  0.33759817 0.2436354  0.32420819
##             address
## age      0.58355417
## debtinc  0.01641204
## creddebt 0.23091537
## income   0.33759817
## othdebt  0.24363540
## employ   0.32420819
## address  1.00000000
```

Cek Kecocokan untuk PCA

```
library(psych)
```

```
## Warning: package 'psych' was built under R version 4.4.2
```

```
## 
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
## 
##     %+%, alpha
```

```
KMO(cor_matrix)
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = cor_matrix)
## Overall MSA =  0.6
## MSA for each item = 
##      age  debtinc creddebt   income  othdebt   employ  address 
##     0.79     0.34     0.64     0.54     0.57     0.89     0.72
```

Data Train dan Data Test

```
set.seed(42)

train_indices <- sample(1:nrow(completed_data1), size = 0.8 * nrow(completed_data1))

train_data <- completed_data1[train_indices, ]  # Training set
test_data <- completed_data1[-train_indices, ]  # Testing set

cat("Training data size: ", nrow(train_data), "\n")
```

```
## Training data size:  560
```

```
cat("Testing data size: ", nrow(test_data), "\n")
```

```
## Testing data size:  140
```

Model Dengan Data Asli sebelum transform

```
logistic_model <- glm(default ~ ., data = train_data, family = binomial)
summary(logistic_model)
```

```
##
## Call:
## glm(formula = default ~ ., family = binomial, data = train_data)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.058314   0.777470  -3.934 8.37e-05 ***
## age          0.039076   0.020209   1.934 0.053167 .
## ed           0.251371   0.149211   1.685 0.092053 .
## employ      -0.190831   0.034035  -5.607 2.06e-08 ***
## address     -0.088847   0.024844  -3.576 0.000349 ***
## income       0.001854   0.014796   0.125 0.900285
## debtinc      0.136713   0.045115   3.030 0.002443 **
## creddebt     0.514725   0.178972   2.876 0.004027 **
## othdebt     -0.061770   0.137542  -0.449 0.653359
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 640.57  on 559  degrees of freedom
## Residual deviance: 463.35  on 551  degrees of freedom
## AIC: 481.35
##
## Number of Fisher Scoring iterations: 5
```

Akurasi

```r
test_predictions_original <- predict(logistic_model, newdata = test_data, type = "response")
threshold <- 0.5
test_class_original <- ifelse(test_predictions_original > threshold, 1, 0)

confusion_matrix_original <- table(Predicted = test_class_original, Actual = test_data$default)
print(confusion_matrix_original)
```

```
##          Actual
## Predicted  0  1
##         0 94 23
##         1  8 15
```

```r
accuracy_original <- mean(test_class_original == test_data$default)
cat("Akurasi Model (Variabel Asli):", accuracy_original, "\n")
```

```
## Akurasi Model (Variabel Asli): 0.7785714
```

Transformasi Data

Transform Min Max Scaling Variabel creddebt

karena creddebt Signifikan dengan koefisien cukup besar dibandingkan variabel lain.

```r
# Fungsi Min-Max Scaling
min_max_scaling <- function(x) {
  (x - min(x, na.rm = TRUE)) / (max(x, na.rm = TRUE) - min(x, na.rm = TRUE))
}

# Terapkan Min-Max Scaling pada variabel creddebt
train_data$creddebt <- min_max_scaling(train_data$creddebt)
test_data$creddebt <- min_max_scaling(test_data$creddebt)

# Periksa hasil scaling
summary(train_data$creddebt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.08748 0.20909 0.32111 0.46885 1.00000
```

Transform Robust Scaling Variabel income

Robust Scaling untuk menangani rentang nilai besar dengan outlier.

```r
train_data$income <- (train_data$income - median(train_data$income)) / IQR(train_data$income)
test_data$income <- (test_data$income - median(train_data$income)) / IQR(train_data$income)
summary(train_data$income)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -0.6942 -0.3636  0.0000  0.2406  0.6364  2.1570
```

Transform Polynomial Variabel age

Alasan: Distribusi sudah cukup normal, tetapi hubungan antara age dan target mungkin non-linear, sehingga menambahkan pangkat kedua

```r
train_data$age <- train_data$age^2
test_data$age <- test_data$age^2
summary(train_data$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     400     841    1156    1283    1600    3364
```

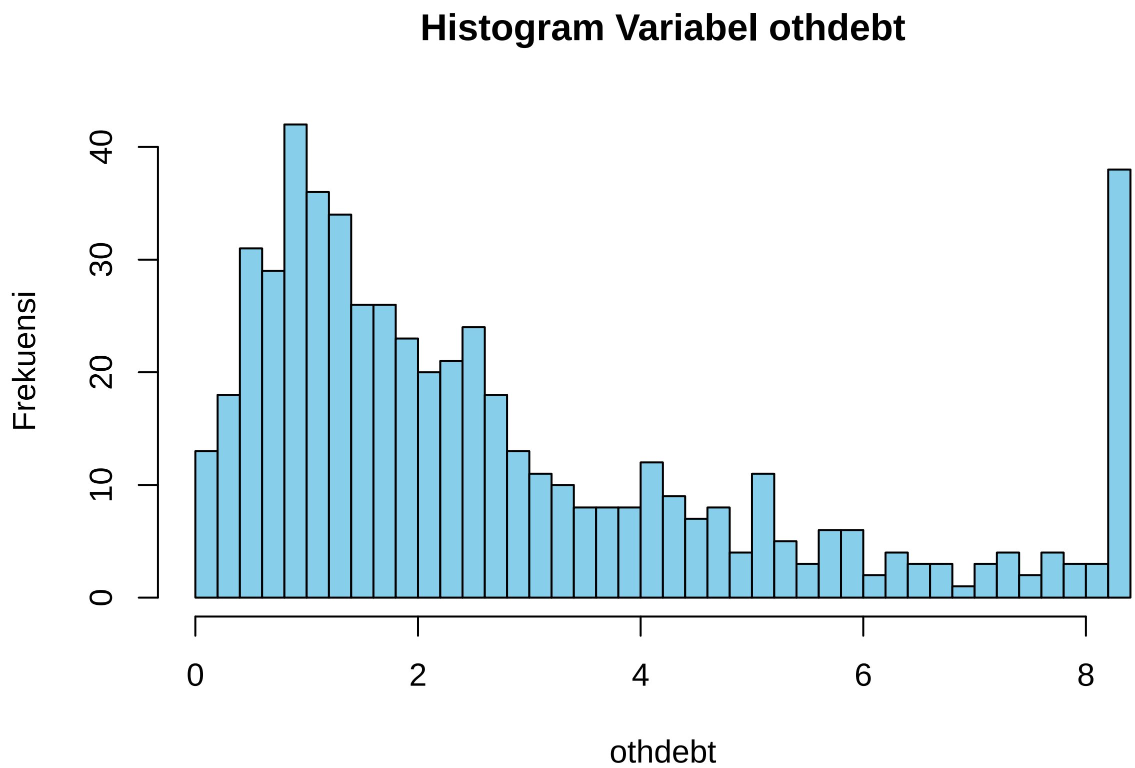
Transform Min Max Variabel age

```r
# Fungsi Min-Max Scaling
min_max_scaling <- function(x) {
  (x - min(x, na.rm = TRUE)) / (max(x, na.rm = TRUE) - min(x, na.rm = TRUE))
}

# Terapkan Min-Max Scaling pada variabel creddebt
train_data$age <- min_max_scaling(train_data$age)
test_data$age <- min_max_scaling(test_data$age)

# Periksa hasil scaling
summary(train_data$age)
```

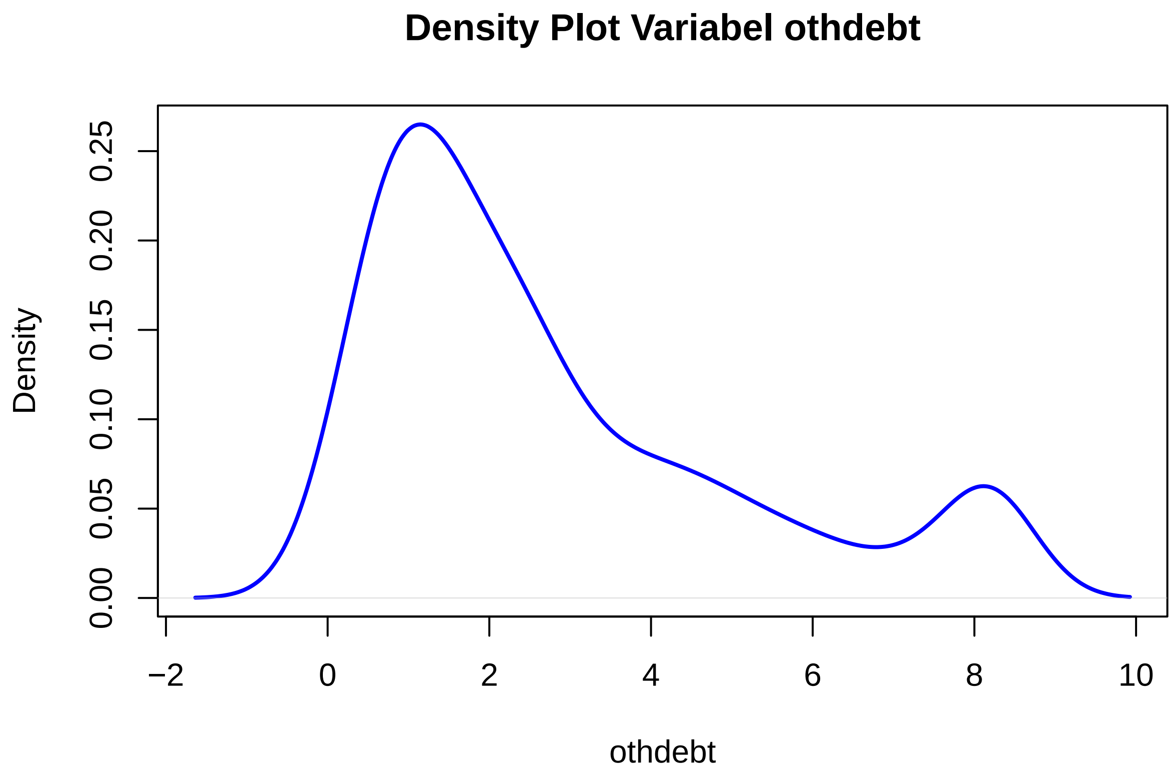
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.1488  0.2551  0.2978  0.4049  1.0000
```

Transform log Variabel othdebt

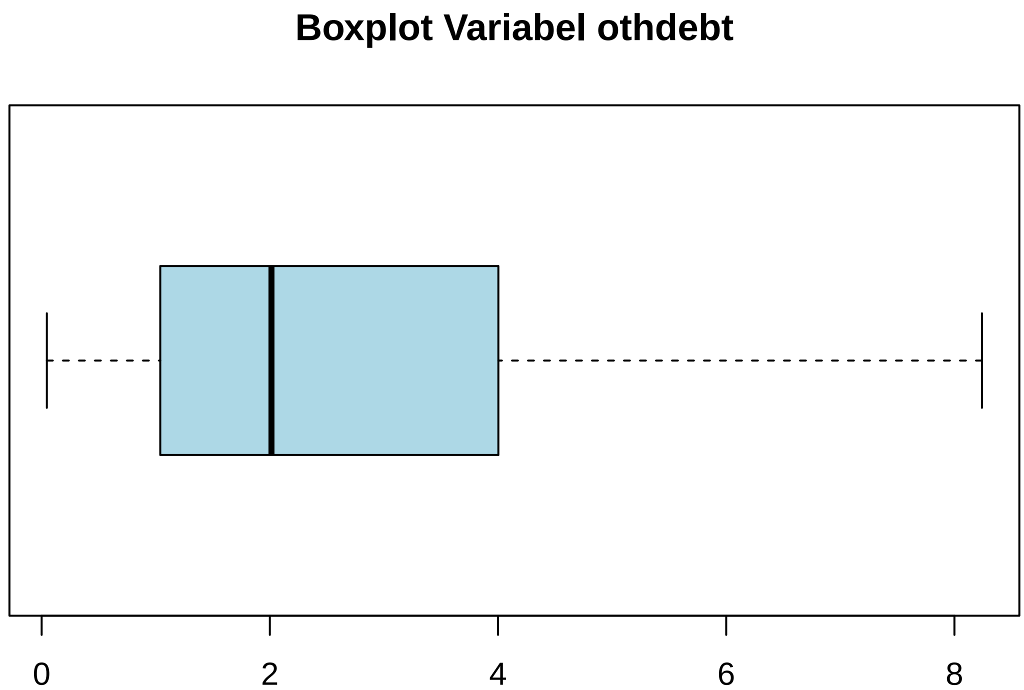
```r
# Histogram untuk memeriksa distribusi othdebt
hist(train_data$othdebt, breaks = 30, col = "skyblue",
     main = "Histogram Variabel othdebt", xlab = "othdebt", ylab = "Frekuensi")
```




```
# Density plot untuk memeriksa distribusi othdebt
plot(density(train_data$othdebt, na.rm = TRUE),
     main = "Density Plot Variabel othdebt",
     xlab = "othdebt", ylab = "Density",
     col = "blue", lwd = 2)
```




```r
# Boxplot untuk mendeteksi outlier
boxplot(train_data$othdebt, main = "Boxplot Variabel othdebt",
        col = "lightblue", horizontal = TRUE)
```




```
train_data$othdebt <- log(train_data$othdebt + 1)
test_data$othdebt <- log(test_data$othdebt + 1)
summary(train_data$othdebt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.04458 0.71307 1.10339 1.16656 1.60981 2.22369
```

ed tidak ditransform

Variabel ed kemungkinan tidak termasuk dalam daftar transformasi karena:

Tidak signifikan dalam model. Distribusi atau tipe datanya tidak relevan untuk transformasi numerik. Korelasi rendah dengan target maupun variabel lainnya.

```
head(train_data)
```

```
##            age  ed employ address     income debtinc    creddebt   othdebt
## 561 0.23245614 2.0     10       4 -0.2975207     3.1 0.004518806 0.5740100
## 321 0.05937922 1.0      8       4 -0.3636364     5.0 0.088290951 0.5979570
## 153 0.30229420 1.0      4      10 -0.3966942    16.3 0.313979658 1.2301825
## 74  0.48886640 3.0      1       5 -0.2644628    10.6 0.359666569 0.8053341
## 228 0.40485830 3.5      5       6  1.3223140     1.9 0.208427643 0.4318311
## 146 0.46018893 2.0      5       3  0.1983471     3.4 0.082054083 0.7122290
##     default
```

```
## 561       0
## 321       0
## 153       0
## 74        0
## 228       0
## 146       0
```

Model Setelah Transformasi

```r
logistic_model1 <- glm(default ~ ., data = train_data, family = binomial)
summary(logistic_model1)
```

```
##
## Call:
## glm(formula = default ~ ., family = binomial, data = train_data)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.64871    0.50098  -3.291 0.000998 ***
## age          1.76523    0.80248   2.200 0.027826 *
## ed           0.28130    0.15059   1.868 0.061763 .
## employ      -0.19037    0.03399  -5.600 2.14e-08 ***
## address     -0.08801    0.02480  -3.548 0.000388 ***
## income       0.53148    0.42020   1.265 0.205939
## debtinc      0.20563    0.04840   4.249 2.15e-05 ***
## creddebt     1.66304    0.73997   2.247 0.024612 *
## othdebt     -1.19315    0.57535  -2.074 0.038101 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 640.57  on 559  degrees of freedom
## Residual deviance: 458.64  on 551  degrees of freedom
## AIC: 476.64
##
## Number of Fisher Scoring iterations: 5
```

Akurasi Setelah Transformasi

```r
test_predictions_original <- predict(logistic_model1, newdata = test_data, type = "response")
threshold <- 0.5
test_class_original <- ifelse(test_predictions_original > threshold, 1, 0)

confusion_matrix_original <- table(Predicted = test_class_original, Actual = test_data$default)
print(confusion_matrix_original)
```

```
##          Actual
## Predicted   0   1
##         1 102  38
```

```
accuracy_original <- mean(test_class_original == test_data$default)
cat("Akurasi Model (Setelah Trnasformasi):", accuracy_original, "\n")
```

```
## Akurasi Model (Setelah Trnasformasi): 0.2714286
```

Feature Selection Berdasarkan Korelasi dengan Target

```
library(caret)

# Chi-Square Test untuk kategori target
chi_sq <- sapply(train_data[, -which(names(train_data) == "default")],
                 function(x) chisq.test(table(x, train_data$default))$p.value)
```

```
## Warning in chisq.test(table(x, train_data$default)): Chi-squared approximation
## may be incorrect
## Warning in chisq.test(table(x, train_data$default)): Chi-squared approximation
## may be incorrect
## Warning in chisq.test(table(x, train_data$default)): Chi-squared approximation
## may be incorrect
## Warning in chisq.test(table(x, train_data$default)): Chi-squared approximation
## may be incorrect
## Warning in chisq.test(table(x, train_data$default)): Chi-squared approximation
## may be incorrect
## Warning in chisq.test(table(x, train_data$default)): Chi-squared approximation
## may be incorrect
## Warning in chisq.test(table(x, train_data$default)): Chi-squared approximation
## may be incorrect
```

```
# Menampilkan p-value untuk setiap variabel
chi_sq
```

```
##          age           ed       employ      address       income      debtinc
## 1.436457e-01 1.134719e-02 6.355054e-06 3.154408e-03 1.279746e-01 1.458427e-05
##     creddebt      othdebt
## 6.420816e-01 6.286732e-01
```

Metode Backward, Forward, dan Best Subset Selection

```
# Membuat model awal dengan semua variabel
full_model <- glm(default ~ ., data = train_data, family = binomial)

# Backward selection menggunakan stepAIC
library(MASS)
backward_model <- stepAIC(full_model, direction = "backward")
```

Backward Selection

```
## Start:  AIC=476.64
## default ~ age + ed + employ + address + income + debtinc + creddebt +
##     othdebt
##
##            Df Deviance    AIC
## - income    1   460.18 476.18
## <none>          458.64 476.64
## - ed        1   462.12 478.12
## - othdebt   1   462.81 478.81
## - age       1   463.37 479.37
## - creddebt  1   463.84 479.84
## - address   1   471.91 487.91
## - debtinc   1   476.73 492.73
## - employ    1   494.72 510.72
##
## Step:  AIC=476.18
## default ~ age + ed + employ + address + debtinc + creddebt +
##     othdebt
##
##            Df Deviance    AIC
## <none>          460.18 476.18
## - othdebt   1   462.92 476.92
## - ed        1   464.68 478.68
## - age       1   465.80 479.80
## - address   1   474.15 488.15
## - creddebt  1   474.58 488.58
## - debtinc   1   486.75 500.75
## - employ    1   495.72 509.72
```

```r
# Melihat model hasil seleksi
summary(backward_model)
```

```
##
## Call:
## glm(formula = default ~ age + ed + employ + address + debtinc +
##     creddebt + othdebt, family = binomial, data = train_data)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.04410    0.39430  -5.184 2.17e-07 ***
## age          1.89344    0.79192   2.391 0.016804 *
## ed           0.31418    0.14792   2.124 0.033672 *
## employ      -0.17653    0.03200  -5.517 3.44e-08 ***
## address     -0.08949    0.02464  -3.632 0.000281 ***
## debtinc      0.16127    0.03287   4.907 9.27e-07 ***
## creddebt     2.22263    0.59688   3.724 0.000196 ***
## othdebt     -0.68525    0.41838  -1.638 0.101451
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 640.57  on 559  degrees of freedom
## Residual deviance: 460.18  on 552  degrees of freedom
```

```
## AIC: 476.18
##
## Number of Fisher Scoring iterations: 5
```

```r
# Membuat model awal dengan intercept saja
null_model <- glm(default ~ 1, data = train_data, family = binomial)

# Full model dengan semua variabel
full_model <- glm(default ~ ., data = train_data, family = binomial)

# Forward selection menggunakan stepAIC
forward_model <- stepAIC(null_model, scope = list(lower = null_model, upper = full_model), direction =
```

Forward Selection

```
## Start:  AIC=642.57
## default ~ 1
##
##            Df Deviance    AIC
## + debtinc   1   550.12 554.12
## + employ    1   595.42 599.42
## + creddebt  1   613.86 617.86
## + othdebt   1   628.90 632.90
## + address   1   629.08 633.08
## + ed        1   630.67 634.67
## + income    1   631.25 635.25
## + age       1   634.79 638.79
## <none>          640.57 642.57
##
## Step:  AIC=554.12
## default ~ debtinc
##
##            Df Deviance    AIC
## + employ    1   495.28 501.28
## + othdebt   1   528.02 534.02
## + address   1   532.23 538.23
## + income    1   538.75 544.75
## + ed        1   539.60 545.60
## + age       1   541.28 547.28
## <none>          550.12 554.12
## + creddebt  1   549.96 555.96
##
## Step:  AIC=501.28
## default ~ debtinc + employ
##
##            Df Deviance    AIC
## + creddebt  1   479.07 487.07
## + income    1   488.82 496.82
## + ed        1   489.52 497.52
## + address   1   490.09 498.09
## <none>          495.28 501.28
```

```
## + age       1   494.43 502.43
## + othdebt   1   495.01 503.01
##
## Step:  AIC=487.07
## default ~ debtinc + employ + creddebt
##
##           Df Deviance    AIC
## + address  1   469.91 479.91
## <none>         479.07 487.07
## + ed       1   477.26 487.26
## + othdebt  1   477.87 487.87
## + income   1   478.96 488.96
## + age      1   479.01 489.01
##
## Step:  AIC=479.91
## default ~ debtinc + employ + creddebt + address
##
##           Df Deviance    AIC
## + age      1   465.65 477.65
## + ed       1   467.24 479.24
## <none>         469.91 479.91
## + income   1   469.36 481.36
## + othdebt  1   469.56 481.56
##
## Step:  AIC=477.65
## default ~ debtinc + employ + creddebt + address + age
##
##           Df Deviance    AIC
## + ed       1   462.92 476.92
## <none>         465.65 477.65
## + othdebt  1   464.68 478.68
## + income   1   465.56 479.56
##
## Step:  AIC=476.92
## default ~ debtinc + employ + creddebt + address + age + ed
##
##           Df Deviance    AIC
## + othdebt  1   460.18 476.18
## <none>         462.92 476.92
## + income   1   462.81 478.81
##
## Step:  AIC=476.18
## default ~ debtinc + employ + creddebt + address + age + ed +
##     othdebt
##
##          Df Deviance    AIC
## <none>        460.18 476.18
## + income  1   458.64 476.64
```

```r
# Melihat model hasil seleksi
summary(forward_model)
```

```
##
## Call:
```

```
## glm(formula = default ~ debtinc + employ + creddebt + address +
##     age + ed + othdebt, family = binomial, data = train_data)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.04410    0.39430  -5.184 2.17e-07 ***
## debtinc      0.16127    0.03287   4.907 9.27e-07 ***
## employ      -0.17653    0.03200  -5.517 3.44e-08 ***
## creddebt     2.22263    0.59688   3.724 0.000196 ***
## address     -0.08949    0.02464  -3.632 0.000281 ***
## age          1.89344    0.79192   2.391 0.016804 *
## ed           0.31418    0.14792   2.124 0.033672 *
## othdebt     -0.68525    0.41838  -1.638 0.101451
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 640.57  on 559  degrees of freedom
## Residual deviance: 460.18  on 552  degrees of freedom
## AIC: 476.18
##
## Number of Fisher Scoring iterations: 5
```

```r
library(leaps)
```

Best Subset Selection

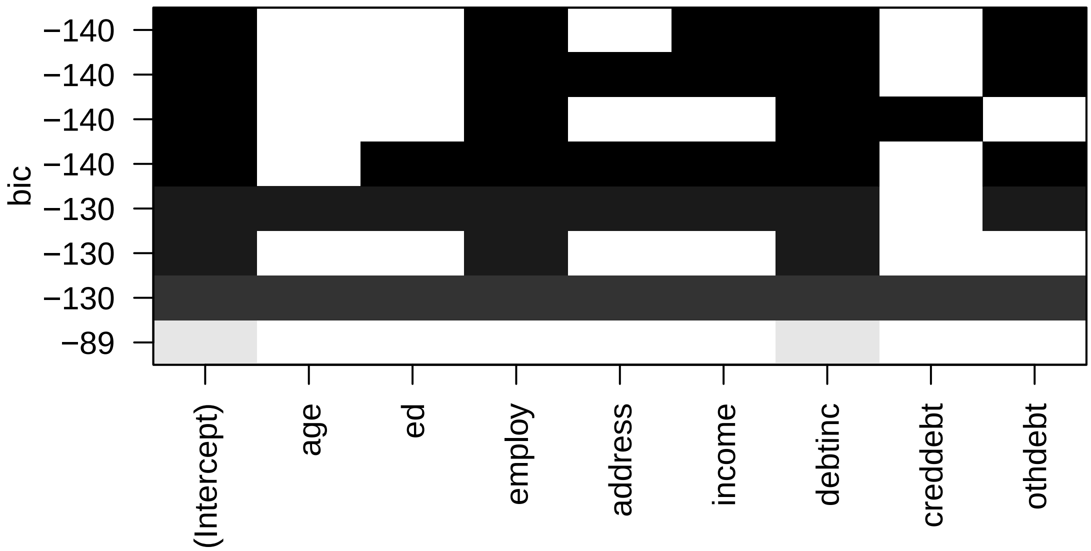
```
## Warning: package 'leaps' was built under R version 4.4.2
```

```r
# Best subset selection
best_subset <- regsubsets(default ~ ., data = train_data, nvmax = 10)  # nvmax: jumlah maksimal variabe

# Menampilkan hasil
summary(best_subset)
```

```
## Subset selection object
## Call: regsubsets.formula(default ~ ., data = train_data, nvmax = 10)
## 8 Variables  (and intercept)
##          Forced in Forced out
## age          FALSE      FALSE
## ed           FALSE      FALSE
## employ       FALSE      FALSE
## address      FALSE      FALSE
## income       FALSE      FALSE
## debtinc      FALSE      FALSE
## creddebt     FALSE      FALSE
## othdebt      FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
```

```
##          age ed  employ address income debtinc creddebt othdebt
## 1  ( 1 ) " " " " " "    " "     " "    "*"     " "      " "
## 2  ( 1 ) " " " " "*"    " "     " "    "*"     " "      " "
## 3  ( 1 ) " " " " "*"    " "     " "    "*"     "*"      " "
## 4  ( 1 ) " " " " "*"    " "     "*"    "*"     " "      "*"
## 5  ( 1 ) " " " " "*"    "*"     "*"    "*"     " "      "*"
## 6  ( 1 ) " " "*" "*"    "*"     "*"    "*"     " "      "*"
## 7  ( 1 ) "*" "*" "*"    "*"     "*"    "*"     " "      "*"
## 8  ( 1 ) "*" "*" "*"    "*"     "*"    "*"     "*"      "*"
```

```
# Plotkan hasil untuk memilih model terbaik
plot(best_subset, scale = "bic")  # Pilihan: "bic", "adjr2", dll.
```



Terbaik

```
logistic_model2 <- glm(default ~ age + ed + employ + address + debtinc + creddebt + othdebt,  data = tra
summary(logistic_model2)
```

```
##
## Call:
## glm(formula = default ~ age + ed + employ + address + debtinc +
##     creddebt + othdebt, family = binomial, data = train_data)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.04410    0.39430  -5.184 2.17e-07 ***
```

```
## age          1.89344    0.79192   2.391 0.016804 *
## ed           0.31418    0.14792   2.124 0.033672 *
## employ      -0.17653    0.03200  -5.517 3.44e-08 ***
## address     -0.08949    0.02464  -3.632 0.000281 ***
## debtinc      0.16127    0.03287   4.907 9.27e-07 ***
## creddebt     2.22263    0.59688   3.724 0.000196 ***
## othdebt     -0.68525    0.41838  -1.638 0.101451
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 640.57  on 559  degrees of freedom
## Residual deviance: 460.18  on 552  degrees of freedom
## AIC: 476.18
##
## Number of Fisher Scoring iterations: 5
```

Akurasi Setelah Forward

```
test_predictions_original <- predict(logistic_model2, newdata = test_data, type = "response")
threshold <- 0.5
test_class_original <- ifelse(test_predictions_original > threshold, 1, 0)

confusion_matrix_original <- table(Predicted = test_class_original, Actual = test_data$default)
print(confusion_matrix_original)
```

```
##          Actual
## Predicted  0  1
##         0 95 21
##         1  7 17
```

```
accuracy_original <- mean(test_class_original == test_data$default)
cat("Akurasi Model (Setelah Forward):", accuracy_original, "\n")
```

```
## Akurasi Model (Setelah Forward): 0.8
```

```
logistic_model3 <- glm(default ~ age + ed + employ + address + debtinc + creddebt, data = train_data, fa
summary(logistic_model3)
```

```
##
## Call:
## glm(formula = default ~ age + ed + employ + address + debtinc +
##     creddebt, family = binomial, data = train_data)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.04281    0.39377  -5.188 2.13e-07 ***
## age          1.62501    0.77567   2.095 0.036175 *
## ed           0.22912    0.13834   1.656 0.097687 .
```

```
## employ      -0.19780    0.02966  -6.669 2.57e-11 ***
## address     -0.09001    0.02451  -3.672 0.000240 ***
## debtinc      0.12300    0.02248   5.471 4.48e-08 ***
## creddebt     2.19088    0.58941   3.717 0.000202 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 640.57  on 559  degrees of freedom
## Residual deviance: 462.92  on 553  degrees of freedom
## AIC: 476.92
##
## Number of Fisher Scoring iterations: 5
```

Akurasi Setelah Forward

```r
test_predictions <- predict(logistic_model3, newdata = test_data, type = "response")
threshold <- 0.5
test_class <- ifelse(test_predictions > threshold, 1, 0)

confusion_matrix_original <- table(Predicted = test_class, Actual = test_data$default)
print(confusion_matrix_original)
```

```
##          Actual
## Predicted  0  1
##         0 93 20
##         1  9 18
```

```r
accuracy_original <- mean(test_class_original == test_data$default)
cat("Akurasi Model (Setelah Forward):", accuracy_original, "\n")
```

```
## Akurasi Model (Setelah Forward): 0.8
```