Data Cleaning

Hazim Fitri

2024-12-21

Contents


```
custdata5 = read.csv('custdata5.csv', sep=';')
custdata5
```

```
##      state.of.res  custid sex is.employed       marital.stat health.ins
## 1       Louisiana   33828   F        TRUE            Married       TRUE
## 2     Connecticut  150055   F        TRUE            Married       TRUE
## 3       Minnesota  204591   F        TRUE Divorced/Separated       TRUE
## 4        New York 1327830   M          NA Divorced/Separated       TRUE
## 5         Montana  950326   M        TRUE      Never Married      FALSE
## 6  North Carolina 1405584   F          NA            Married       TRUE
## 7        Maryland 1364927   M        TRUE            Married       TRUE
## 8            Ohio 1031473   M        TRUE            Widowed       TRUE
## 9    Pennsylvania 1342272   F       FALSE            Married      FALSE
## 10        Montana  699499   M        TRUE      Never Married       TRUE
## 11   Pennsylvania   84879   F        TRUE            Married       TRUE
## 12      Minnesota  938913   M        TRUE            Married       TRUE
## 13        Florida  220135   M        TRUE Divorced/Separated       TRUE
## 14     California 1239370   F        TRUE            Married       TRUE
## 15       Illinois  589826   F        TRUE      Never Married       TRUE
## 16           Ohio 1037857   M        TRUE            Married       TRUE
## 17 North Carolina 1263543   M          NA            Married       TRUE
## 18     Washington 1058755   F        TRUE Divorced/Separated       TRUE
## 19       Michigan  120705   M        TRUE      Never Married       TRUE
## 20     California  431729   M       FALSE            Married       TRUE
## 21      Minnesota 1160167   M       FALSE            Married       TRUE
## 22      Louisiana  863276   F       FALSE            Married       TRUE
## 23          Texas  445937   F          NA            Widowed       TRUE
## 24       New York 1185185   M        TRUE            Married       TRUE
## 25           Ohio 1024809   M        TRUE            Married       TRUE
## 26          Texas 1359848   F       FALSE            Married      FALSE
## 27       Maryland 1114096   M        TRUE            Married       TRUE
## 28       Illinois  532971   M        TRUE            Married       TRUE
## 29  West Virginia  952729   F        TRUE      Never Married       TRUE
## 30           Iowa 1236675   M        TRUE Divorced/Separated       TRUE
## 31   Pennsylvania  518899   M        TRUE            Married       TRUE
## 32        Indiana 1283783   F        TRUE      Never Married       TRUE
## 33           Ohio 1038537   F        TRUE      Never Married      FALSE
```

```
## 34        Georgia  774920   M        TRUE      Never Married       TRUE
## 35         Nevada  867842   F       FALSE            Widowed       TRUE
## 36       Nebraska  474507   F        TRUE            Married       TRUE
## 37       New York 1361012   F       FALSE            Widowed       TRUE
## 38       Illinois  863424   F        TRUE      Never Married       TRUE
## 39        Georgia  679364   F       FALSE            Married       TRUE
## 40      Tennessee 1370515   F          NA            Married       TRUE
##                    housing.type recent.move num.vehicles age is.employed.fix1
## 1  Homeowner with mortgage/loan       FALSE            2  49         employed
## 2                        Rented       FALSE            2  24         employed
## 3                        Rented       FALSE            3  49         employed
## 4  Homeowner with mortgage/loan       FALSE            1  67          missing
## 5                        Rented       FALSE            1  26         employed
## 6  Homeowner with mortgage/loan       FALSE            2  55          missing
## 7  Homeowner with mortgage/loan       FALSE           NA  59         employed
## 8  Homeowner with mortgage/loan       FALSE            4  65         employed
## 9                        Rented       FALSE            2  27     not employed
## 10                       Rented       FALSE            2  29         employed
## 11     Homeowner free and clear       FALSE            2  42         employed
## 12     Homeowner free and clear       FALSE            1  61          missing
## 13                       Rented        TRUE            2  64         employed
## 14 Homeowner with mortgage/loan        TRUE            2  50         employed
## 15                       Rented       FALSE            1  NA         employed
## 16 Homeowner with mortgage/loan       FALSE            3  22         employed
## 17     Homeowner free and clear       FALSE            3  77          missing
## 18     Homeowner free and clear       FALSE            1  59         employed
## 19                       Rented        TRUE            0  30         employed
## 20 Homeowner with mortgage/loan       FALSE            2  75          missing
## 21 Homeowner with mortgage/loan       FALSE            2  74          missing
## 22                       Rented       FALSE            1  46     not employed
## 23                       Rented       FALSE            1  86          missing
## 24 Homeowner with mortgage/loan       FALSE           NA  71          missing
## 25 Homeowner with mortgage/loan        TRUE            1  24         employed
## 26     Homeowner free and clear       FALSE            4  59     not employed
## 27                       Rented       FALSE            1  NA         employed
## 28                       Rented       FALSE            2  33         employed
## 29                       Rented       FALSE            1  28         employed
## 30 Homeowner with mortgage/loan       FALSE            1  39         employed
## 31 Homeowner with mortgage/loan       FALSE            3  49         employed
## 32 Homeowner with mortgage/loan       FALSE            1  63         employed
## 33     Homeowner free and clear       FALSE            1  42         employed
## 34 Homeowner with mortgage/loan       FALSE            4  43         employed
## 35     Homeowner free and clear       FALSE            1  75          missing
## 36 Homeowner with mortgage/loan       FALSE            3  52         employed
## 37     Homeowner free and clear       FALSE            1  NA          missing
## 38                       Rented       FALSE           NA  42         employed
## 39     Homeowner free and clear       FALSE            2  55     not employed
## 40 Homeowner with mortgage/loan       FALSE            2  62          missing
##    age.range Yearly.Income
## 1    (25,65]         48000
## 2     [0,25]         63000
## 3    (25,65]         20000
## 4   (65,Inf]            NA
## 5    (25,65]         34200
```

```
## 6    (25,65]         80000
## 7    (25,65]         90000
## 8    (25,65]         22800
## 9    (25,65]         33500
## 10   (25,65]        100600
## 11   (25,65]         25000
## 12   (25,65]         63000
## 13   (25,65]            NA
## 14   (25,65]         75000
## 15   (25,65]         44600
## 16    [0,25]         20000
## 17  (65,Inf]         32000
## 18   (25,65]         24800
## 19   (25,65]        120000
## 20  (65,Inf]         63000
## 21  (65,Inf]         61000
## 22   (25,65]         80000
## 23  (65,Inf]        165200
## 24  (65,Inf]            NA
## 25    [0,25]         90000
## 26   (25,65]         55000
## 27   (25,65]         71800
## 28   (25,65]            NA
## 29   (25,65]        402000
## 30   (25,65]         14000
## 31   (25,65]         22800
## 32   (25,65]         93500
## 33   (25,65]            NA
## 34   (25,65]        110000
## 35  (65,Inf]         71800
## 36   (25,65]         38080
## 37  (65,Inf]         40000
## 38   (25,65]            NA
## 39   (25,65]         42000
## 40   (25,65]        104000
```

```
str(custdata5)
```

Terangkan berkenaan latar belakang data dan statistik ringkas data

```
## 'data.frame':    40 obs. of  13 variables:
##  $ state.of.res    : chr  "Louisiana" "Connecticut" "Minnesota" "New York" ...
##  $ custid          : int  33828 150055 204591 1327830 950326 1405584 1364927 1031473 1342272 699499
##  $ sex             : chr  "F" "F" "F" "M" ...
##  $ is.employed     : logi  TRUE TRUE TRUE NA TRUE NA ...
##  $ marital.stat    : chr  "Married" "Married" "Divorced/Separated" "Divorced/Separated" ...
##  $ health.ins      : logi  TRUE TRUE TRUE TRUE FALSE TRUE ...
##  $ housing.type    : chr  "Homeowner with mortgage/loan" "Rented" "Rented" "Homeowner with mortgage/l
##  $ recent.move     : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ num.vehicles    : int  2 2 3 1 1 2 NA 4 2 2 ...
##  $ age             : int  49 24 49 67 26 55 59 65 27 29 ...
```

```
##  $ is.employed.fix1: chr  "employed" "employed" "employed" "missing" ...
##  $ age.range       : chr  "(25,65]" "[0,25]" "(25,65]" "(65,Inf]" ...
##  $ Yearly.Income   : int  48000 63000 20000 NA 34200 80000 90000 22800 33500 100600 ...
```

```
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
```

```
data = custdata5 %>%
  mutate(sex=as.factor(sex)) %>%
  mutate(marital.stat=as.factor(marital.stat)) %>%
  mutate(health.ins=as.factor(health.ins)) %>%
  mutate(housing.type=as.factor(housing.type)) %>%
  mutate(recent.move=as.factor(recent.move)) %>%
  mutate(is.employed.fix1=as.factor(is.employed.fix1))
```