GGplot_Bonus_Rui_Peng.R

raypeng

2025-04-06

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
PATH = "/Users/raypeng/Documents/IS 5213 Data science and big data/HMEQ_Scrubbed"
FILE_NAME = "HMEQ_Scrubbed.csv"

INFILE = paste(PATH, FILE_NAME, sep = "/")

setwd(PATH)
df = read.csv(FILE_NAME)

str(df)
```

```
## 'data.frame':    5960 obs. of  29 variables:
##  $ TARGET_BAD_FLAG    : int  1 1 1 1 0 1 1 1 1 1 ...
##  $ TARGET_LOSS_AMT    : int  641 1109 767 1425 0 335 1841 373 1217 1523 ...
##  $ LOAN               : int  1100 1300 1500 1500 1700 1700 1800 1800 2000 2000 ...
##  $ IMP_MORTDUE        : num  25860 70053 13500 65000 97800 ...
##  $ M_MORTDUE          : int  0 0 0 1 0 0 0 0 0 1 ...
##  $ IMP_VALUE          : num  39025 68400 16700 89000 112000 ...
##  $ M_VALUE            : int  0 0 0 1 0 0 0 0 0 0 ...
##  $ IMP_YOJ            : num  10.5 7 4 7 3 9 5 11 3 16 ...
##  $ M_YOJ              : int  0 0 0 1 0 0 0 0 0 0 ...
##  $ IMP_DEROG          : int  0 0 0 1 0 0 3 0 0 0 ...
##  $ M_DEROG            : int  0 0 0 1 0 0 0 0 0 0 ...
##  $ IMP_DELINQ         : int  0 2 0 1 0 0 2 0 2 0 ...
##  $ M_DELINQ           : int  0 0 0 1 0 0 0 0 0 0 ...
##  $ IMP_CLAGE          : num  94.4 121.8 149.5 174 93.3 ...
##  $ M_CLAGE            : int  0 0 0 1 0 0 0 0 0 0 ...
##  $ IMP_NINQ           : int  1 0 1 1 0 1 1 0 1 0 ...
##  $ M_NINQ             : int  0 0 0 1 0 0 0 0 0 0 ...
```

```
##  $ IMP_CLNO           : int  9 14 10 20 14 8 17 8 12 13 ...
##  $ M_CLNO             : int  0 0 0 1 0 0 0 0 0 0 ...
##  $ IMP_DEBTINC        : num  35 35 35 35 35 ...
##  $ M_DEBTINC          : int  1 1 1 1 1 0 1 0 1 1 ...
##  $ FLAG.Job.Mgr       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ FLAG.Job.Office    : int  0 0 0 0 1 0 0 0 0 0 ...
##  $ FLAG.Job.Other     : int  1 1 1 0 0 1 1 1 1 0 ...
##  $ FLAG.Job.ProfExe   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ FLAG.Job.Sales     : int  0 0 0 0 0 0 0 0 0 1 ...
##  $ FLAG.Job.Self      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ FLAG.Reason.DebtCon: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ FLAG.Reason.HomeImp: int  1 1 1 0 1 1 1 1 1 1 ...
```

```
summary(df)
```

```
##  TARGET_BAD_FLAG  TARGET_LOSS_AMT      LOAN        IMP_MORTDUE
##  Min.   :0.0000   Min.   :    0   Min.   : 1100   Min.   :  2063
##  1st Qu.:0.0000   1st Qu.:    0   1st Qu.:11100   1st Qu.: 48139
##  Median :0.0000   Median :    0   Median :16300   Median : 65000
##  Mean   :0.1995   Mean   : 2676   Mean   :18608   Mean   : 72999
##  3rd Qu.:0.0000   3rd Qu.:    0   3rd Qu.:23300   3rd Qu.: 88200
##  Max.   :1.0000   Max.   :78987   Max.   :89900   Max.   :399550
##    M_MORTDUE         IMP_VALUE         M_VALUE           IMP_YOJ
##  Min.   :0.00000   Min.   :  8000   Min.   :0.00000   Min.   : 0.000
##  1st Qu.:0.00000   1st Qu.: 66490   1st Qu.:0.00000   1st Qu.: 3.000
##  Median :0.00000   Median : 89000   Median :0.00000   Median : 7.000
##  Mean   :0.08691   Mean   :101536   Mean   :0.01879   Mean   : 8.756
##  3rd Qu.:0.00000   3rd Qu.:119005   3rd Qu.:0.00000   3rd Qu.:12.000
##  Max.   :1.00000   Max.   :855909   Max.   :1.00000   Max.   :41.000
##      M_YOJ           IMP_DEROG          M_DEROG         IMP_DELINQ
##  Min.   :0.00000   Min.   : 0.0000   Min.   :0.0000   Min.   : 0.000
##  1st Qu.:0.00000   1st Qu.: 0.0000   1st Qu.:0.0000   1st Qu.: 0.000
##  Median :0.00000   Median : 0.0000   Median :0.0000   Median : 0.000
##  Mean   :0.08641   Mean   : 0.3431   Mean   :0.1188   Mean   : 0.503
##  3rd Qu.:0.00000   3rd Qu.: 0.0000   3rd Qu.:0.0000   3rd Qu.: 1.000
##  Max.   :1.00000   Max.   :10.0000   Max.   :1.0000   Max.   :15.000
##     M_DELINQ         IMP_CLAGE         M_CLAGE           IMP_NINQ
##  Min.   :0.00000   Min.   :   0.0   Min.   :0.00000   Min.   : 0.00
##  1st Qu.:0.00000   1st Qu.: 117.4   1st Qu.:0.00000   1st Qu.: 0.00
##  Median :0.00000   Median : 174.0   Median :0.00000   Median : 1.00
##  Mean   :0.09732   Mean   : 179.5   Mean   :0.05168   Mean   : 1.17
##  3rd Qu.:0.00000   3rd Qu.: 227.1   3rd Qu.:0.00000   3rd Qu.: 2.00
##  Max.   :1.00000   Max.   :1168.2   Max.   :1.00000   Max.   :17.00
##      M_NINQ           IMP_CLNO         M_CLNO         IMP_DEBTINC
##  Min.   :0.00000   Min.   : 0.00   Min.   :0.00000   Min.   :  0.5245
##  1st Qu.:0.00000   1st Qu.:15.00   1st Qu.:0.00000   1st Qu.: 30.7632
##  Median :0.00000   Median :20.00   Median :0.00000   Median : 35.0000
##  Mean   :0.08557   Mean   :21.25   Mean   :0.03725   Mean   : 34.0393
##  3rd Qu.:0.00000   3rd Qu.:26.00   3rd Qu.:0.00000   3rd Qu.: 37.9499
##  Max.   :1.00000   Max.   :71.00   Max.   :1.00000   Max.   :203.3122
##    M_DEBTINC       FLAG.Job.Mgr    FLAG.Job.Office  FLAG.Job.Other
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000
```

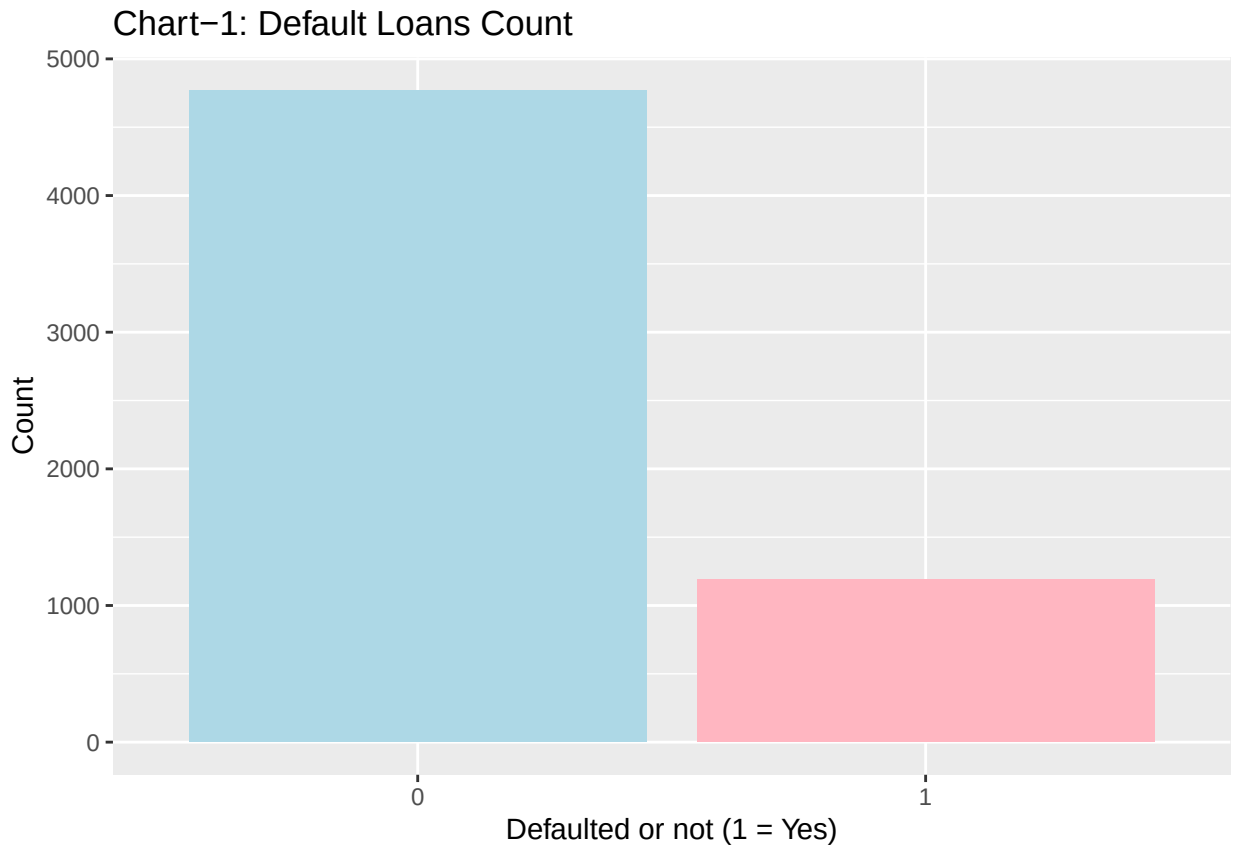
```
##  Mean   :0.2126   Mean   :0.1287   Mean   :0.1591   Mean   :0.4007
##  3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##  FLAG.Job.ProfExe FLAG.Job.Sales    FLAG.Job.Self     FLAG.Reason.DebtCon
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.0000
##  Median :0.0000   Median :0.00000   Median :0.00000   Median :1.0000
##  Mean   :0.2141   Mean   :0.01829   Mean   :0.03238   Mean   :0.6591
##  3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.00000   Max.   :1.0000
##  FLAG.Reason.HomeImp
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.2987
##  3rd Qu.:1.0000
##  Max.   :1.0000
```

```
head(df)
```

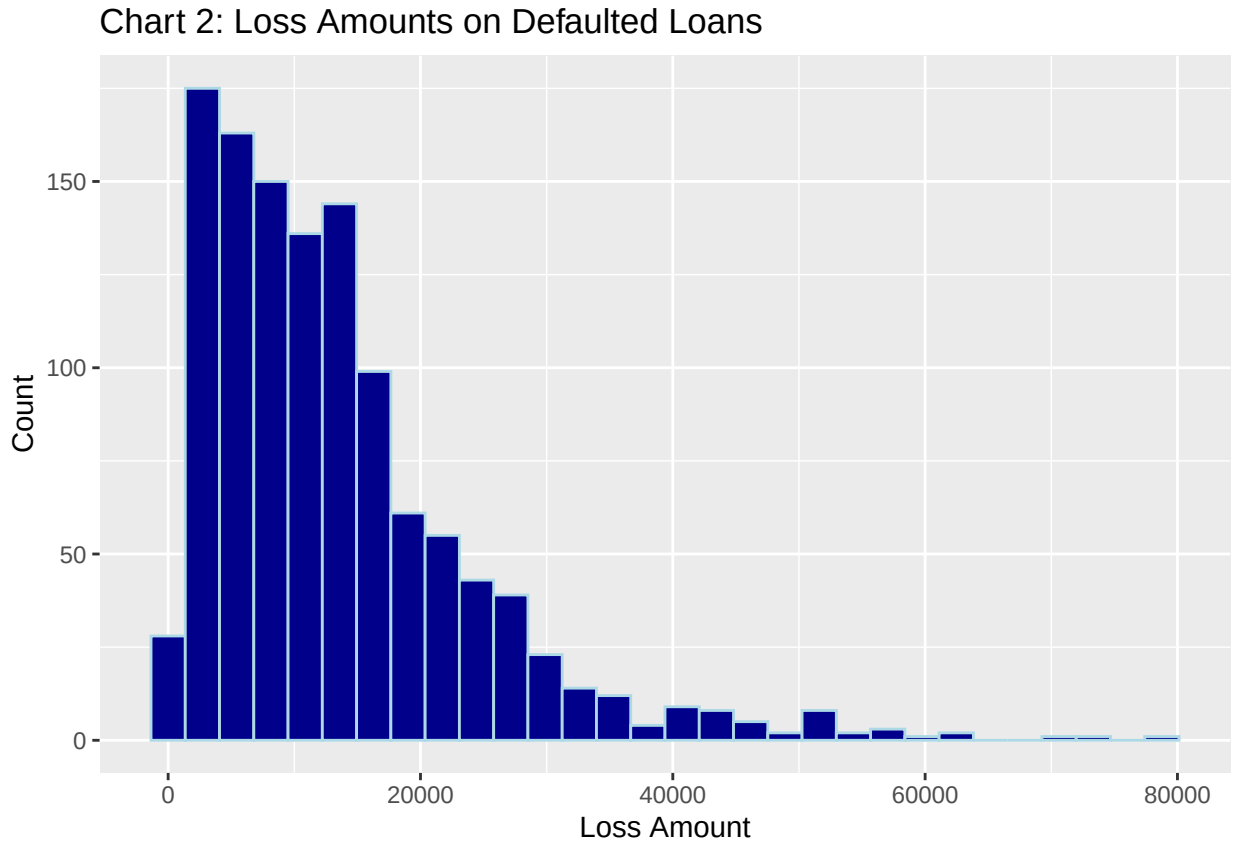
```
##   TARGET_BAD_FLAG TARGET_LOSS_AMT LOAN IMP_MORTDUE M_MORTDUE IMP_VALUE M_VALUE
## 1               1             641 1100       25860         0     39025       0
## 2               1            1109 1300       70053         0     68400       0
## 3               1             767 1500       13500         0     16700       0
## 4               1            1425 1500       65000         1     89000       1
## 5               0               0 1700       97800         0    112000       0
## 6               1             335 1700       30548         0     40320       0
##   IMP_YOJ M_YOJ IMP_DEROG M_DEROG IMP_DELINQ M_DELINQ IMP_CLAGE M_CLAGE
## 1    10.5     0         0       0          0        0  94.36667       0
## 2     7.0     0         0       0          2        0 121.83333       0
## 3     4.0     0         0       0          0        0 149.46667       0
## 4     7.0     1         1       1          1        1 174.00000       1
## 5     3.0     0         0       0          0        0  93.33333       0
## 6     9.0     0         0       0          0        0 101.46600       0
##   IMP_NINQ M_NINQ IMP_CLNO M_CLNO IMP_DEBTINC M_DEBTINC FLAG.Job.Mgr
## 1        1      0        9      0    35.00000         1            0
## 2        0      0       14      0    35.00000         1            0
## 3        1      0       10      0    35.00000         1            0
## 4        1      1       20      1    35.00000         1            0
## 5        0      0       14      0    35.00000         1            0
## 6        1      0        8      0    37.11361         0            0
##   FLAG.Job.Office FLAG.Job.Other FLAG.Job.ProfExe FLAG.Job.Sales FLAG.Job.Self
## 1               0              1                0              0            0
## 2               0              1                0              0            0
## 3               0              1                0              0            0
## 4               0              0                0              0            0
## 5               1              0                0              0            0
## 6               0              1                0              0            0
##   FLAG.Reason.DebtCon FLAG.Reason.HomeImp
## 1                   0                   1
## 2                   0                   1
## 3                   0                   1
## 4                   0                   0
## 5                   0                   1
```

```
## 6                    0                    1
```

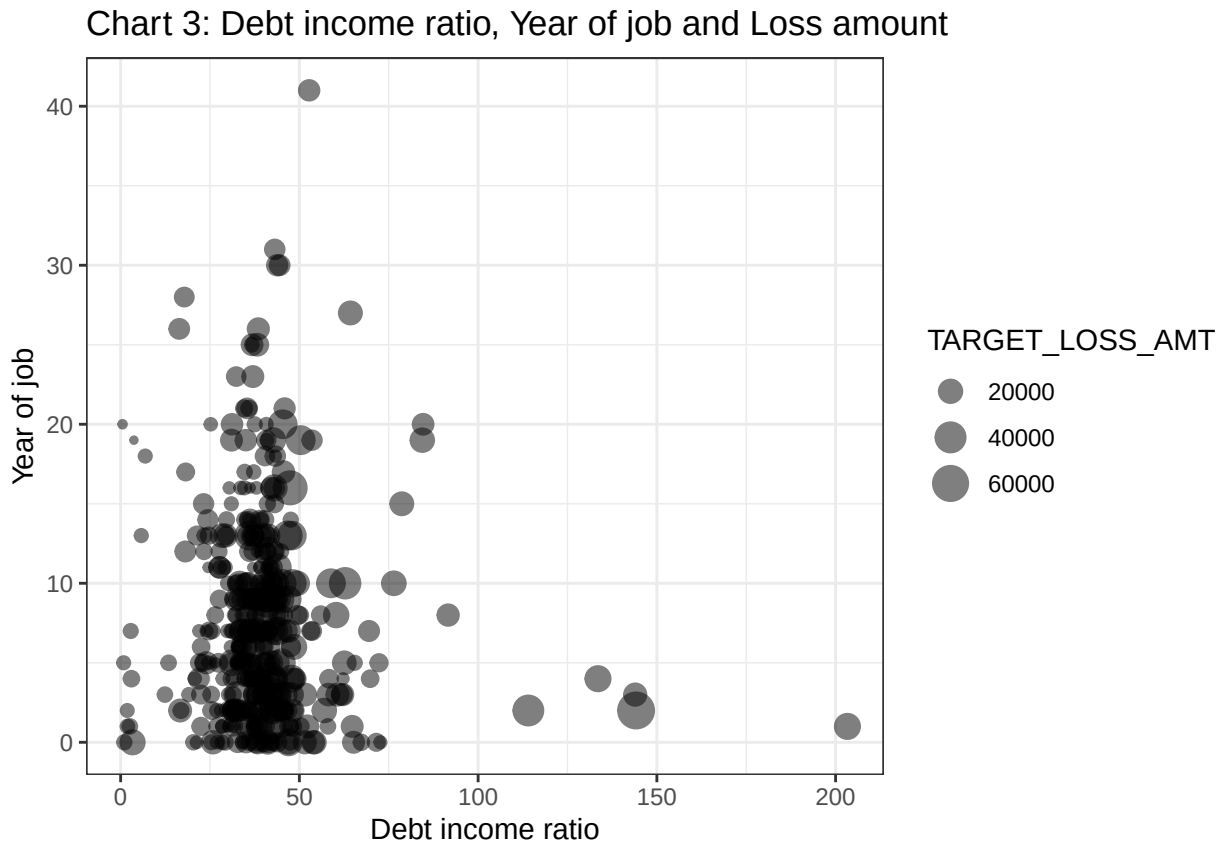
```
#Chart 1
#Defaulted loans are about 1/5 among all the data.
df %>%
  ggplot( aes(x = factor(TARGET_BAD_FLAG)) )+
  geom_bar( fill = c("lightblue", "lightpink") ) +
  labs(title = "Chart-1: Default Loans Count",
       x = "Defaulted or not (1 = Yes)",
       y = "Count")
```



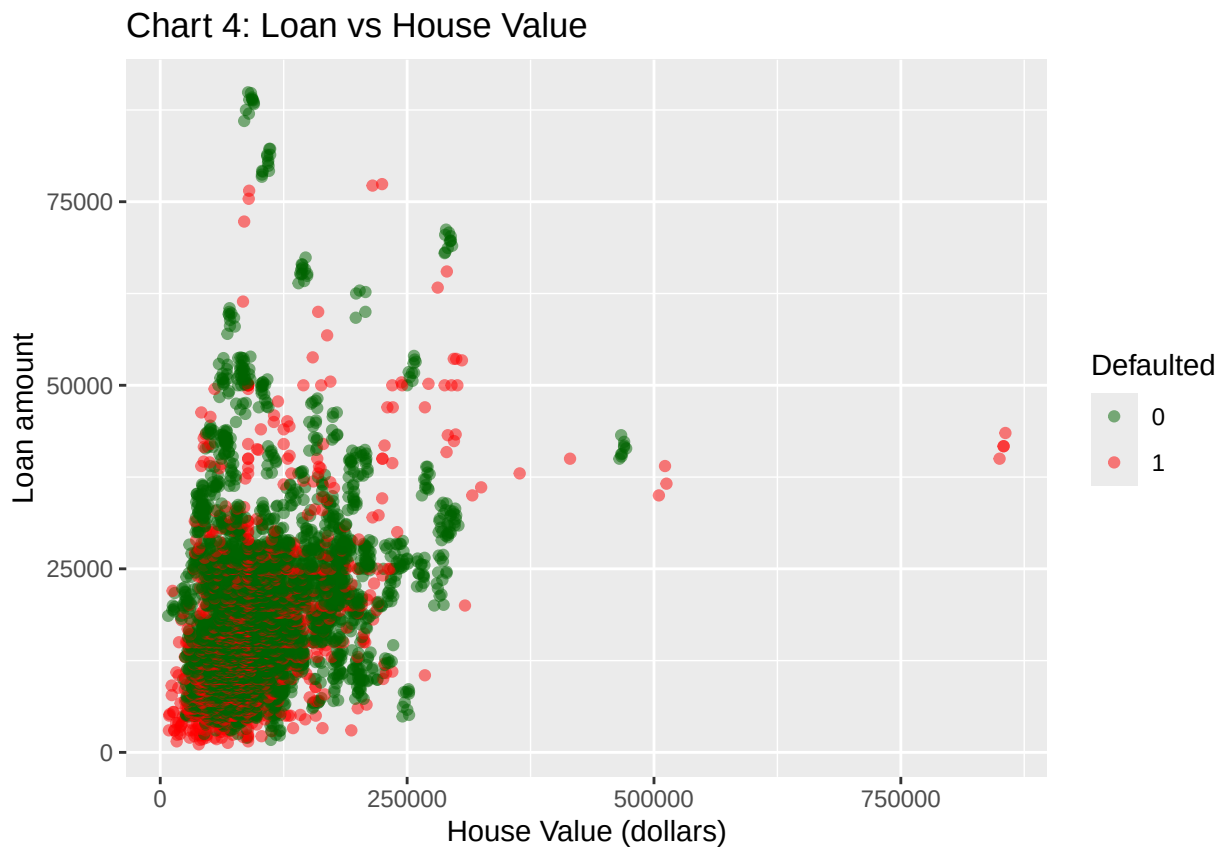

```
#Chart 2
#Loss amounts median and mean are around 10,000 dollars.
df %>%
  filter(TARGET_BAD_FLAG == 1) %>%
  ggplot( aes(x = TARGET_LOSS_AMT) )+
  geom_histogram( fill = "darkblue", col = "lightblue", bins = 30 ) +
  labs(title = "Chart 2: Loss Amounts on Defaulted Loans",
       x = "Loss Amount",
       y = "Count")
```



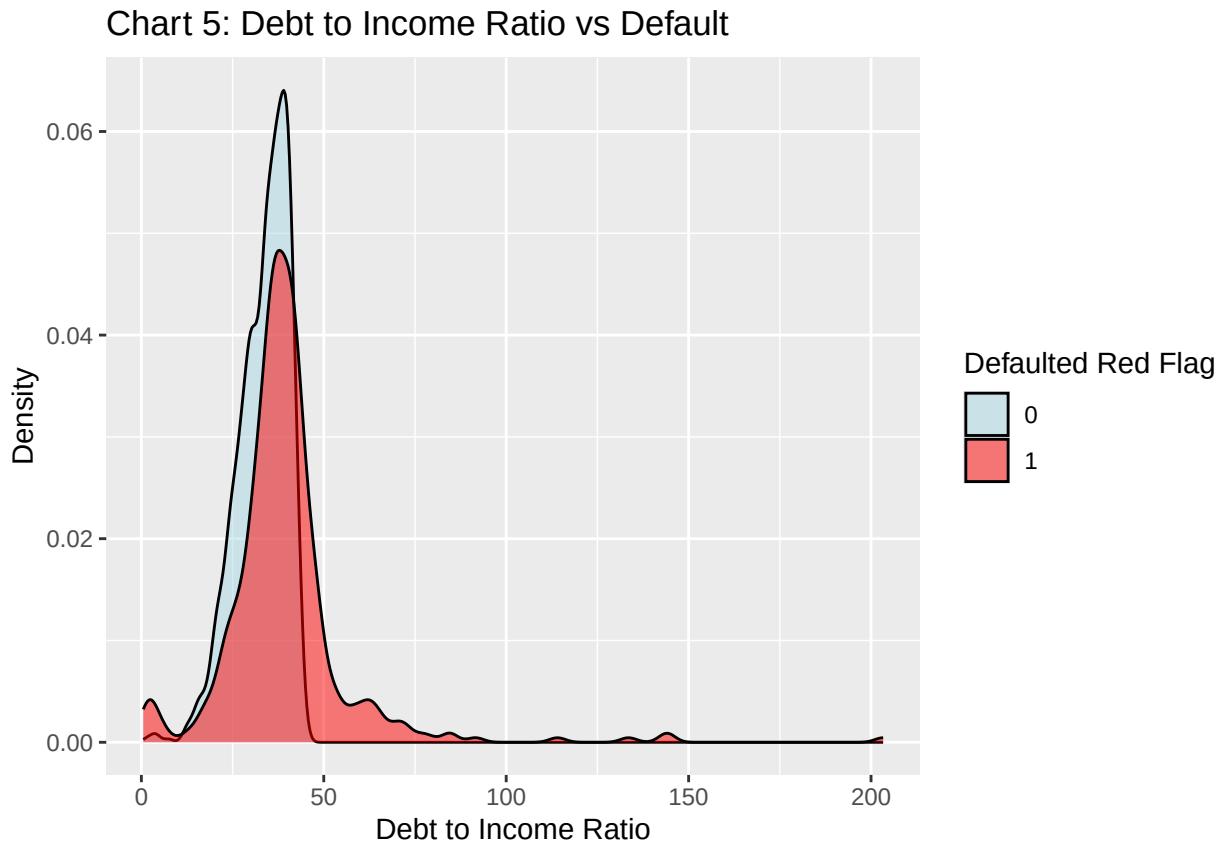

```
#Chart 3
#People with less job experience tend to have bad loans.
#Higher debt-inocme ration may cause larger amounts of defaulted loan.
df %>%
  filter( TARGET_BAD_FLAG == 1 & M_DEBTINC == 0 ) %>%
  ggplot( aes(IMP_DEBTINC, IMP_YOJ))+
  geom_point(alpha = 0.5,
             aes(size = TARGET_LOSS_AMT))+
  theme_bw()+
  labs(title = "Chart 3: Debt income ratio, Year of job and Loss amount",
       x = "Debt income ratio",
       y = "Year of job")
```



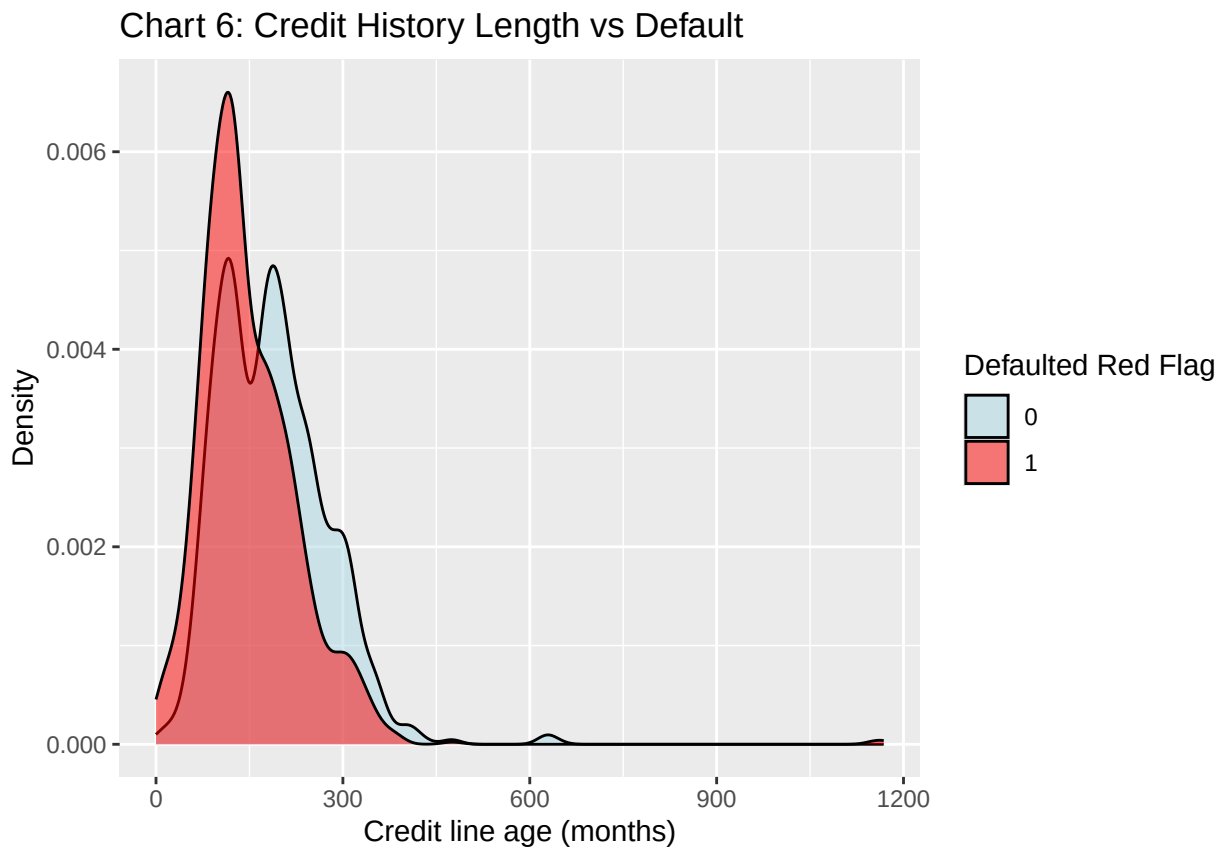

```
#Chart 4
#People with less loan amount tend to default.
df %>%
  ggplot( aes(x = IMP_VALUE,
              y = LOAN,
              color = factor(TARGET_BAD_FLAG)))+
  geom_point(alpha = 0.5)+
  scale_color_manual(values = c("0" = "darkgreen", "1" = "red")) +
  labs(title = "Chart 4: Loan vs House Value",
       color = "Defaulted",
       x = "House Value (dollars)",
       y = "Loan amount")
```




```
#Chart 5
#People with higher Debt-income ratio tend to default.
df %>%
  filter(M_DEBTINC == 0) %>%
  ggplot( aes(x = IMP_DEBTINC,
              fill = factor(TARGET_BAD_FLAG)) )+
  geom_density(alpha = 0.5) +
  scale_fill_manual(values = c("0" = "lightblue", "1" = "red"))+
  labs(title = "Chart 5: Debt to Income Ratio vs Default",
       x = "Debt to Income Ratio",
       y = "Density",
       fill = "Defaulted Red Flag")
```




```
#Chart 6
#People with shorter credit line age tend to default on loans.
df %>%
  filter(M_CLAGE == 0) %>%
  ggplot( aes(x = IMP_CLAGE,
              fill = factor(TARGET_BAD_FLAG))) +
  geom_density(alpha = 0.5) +
  scale_fill_manual(values = c("0" = "lightblue", "1" = "red"))+
  labs(title = "Chart 6: Credit History Length vs Default",
       x = "Credit line age (months)",
       y = "Density",
       fill = "Defaulted Red Flag")
```

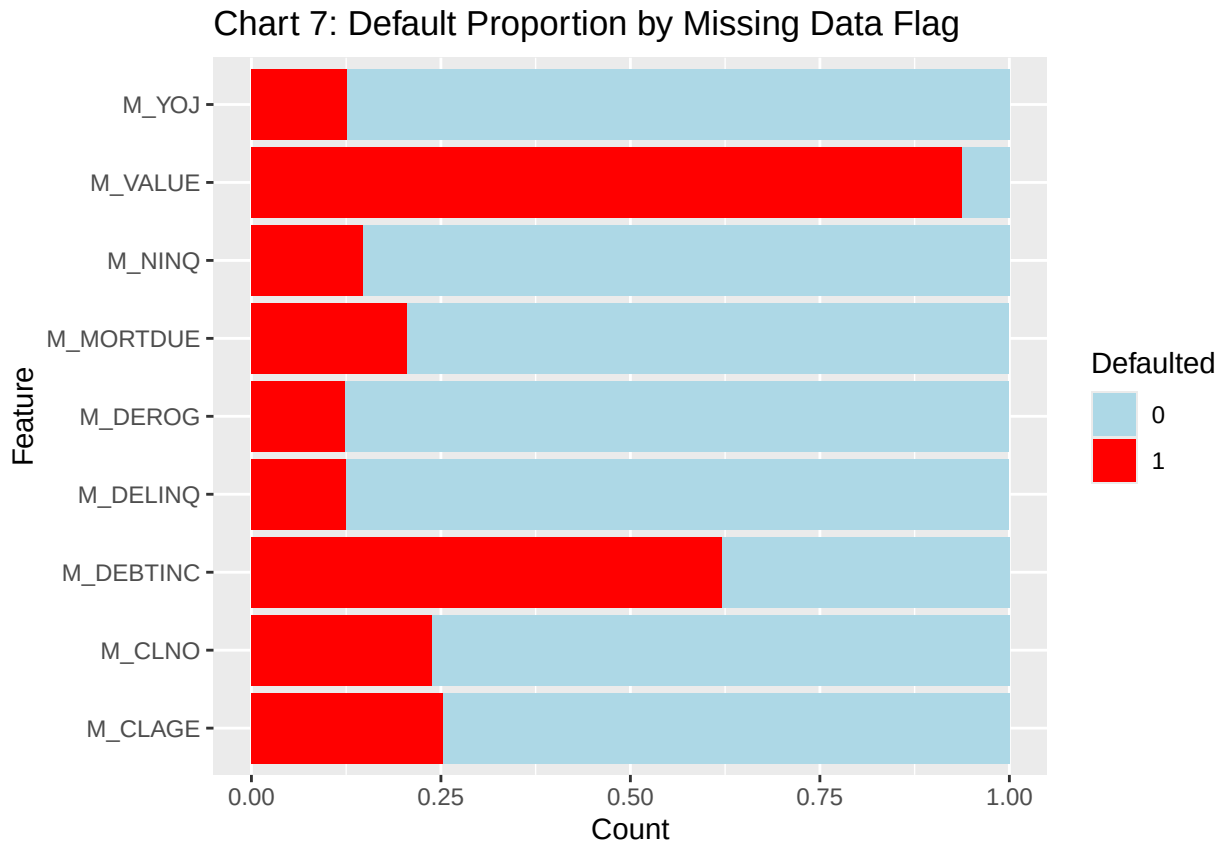



```
#Chart 7
#In this dataset, if the person's house value and debt-income ratio data is missing, then
#this person may default on the loan.

#Chart 8
#In those people whose debt-income ratio data is missing,
#over half of them will default on the loan.
miss_cols = c("M_MORTDUE", "M_VALUE", "M_YOJ", "M_DEROG", "M_DELINQ",
              "M_CLAGE", "M_NINQ", "M_CLNO", "M_DEBTINC")
df_miss = df %>%
  select(TARGET_BAD_FLAG, all_of(miss_cols)) %>%
  pivot_longer(-TARGET_BAD_FLAG, names_to = "Feature", values_to = "Missing") %>%
  filter(Missing == 1)

p1 = ggplot(df_miss, aes(x = Feature,
                    fill = factor(TARGET_BAD_FLAG)))+
  geom_bar(position = "fill")+
#  geom_bar(alpha = 0.8)+
  scale_fill_manual(values = c("0" = "lightblue", "1" = "red"))+
  labs(title = "Chart 7: Default Proportion by Missing Data Flag",
       y = "Count",
       fill = "Defaulted")+
  coord_flip()

print(p1)
```




```
p2 = ggplot(df_miss, aes(x = Feature,
                         fill = factor(TARGET_BAD_FLAG)))+
#  geom_bar(position = "fill")+
  geom_bar(alpha = 0.8)+
  scale_fill_manual(values = c("0" = "lightblue", "1" = "red"))+
  labs(title = "Chart 8: Default Count by Missing Data Feature",
       y = "Count",
       fill = "Defaulted")+
  coord_flip()

print(p2)
```

Chart 8: Default Count by Missing Data Feature