BONUS Missing Value Imputation

Zaremba, Dan

R markdown is a plain-text file format for integrating text and R code, and creating transparent, reproducible and interactive reports. An R markdown file (.Rmd) contains metadata, markdown and R code “chunks,”” and can be “knit” into numerous output types. Answer the test questions by adding R code to the fenced code areas below each item. There are questions that require a written answer that also need to be answered. Enter your comments in the space provided as shown below:

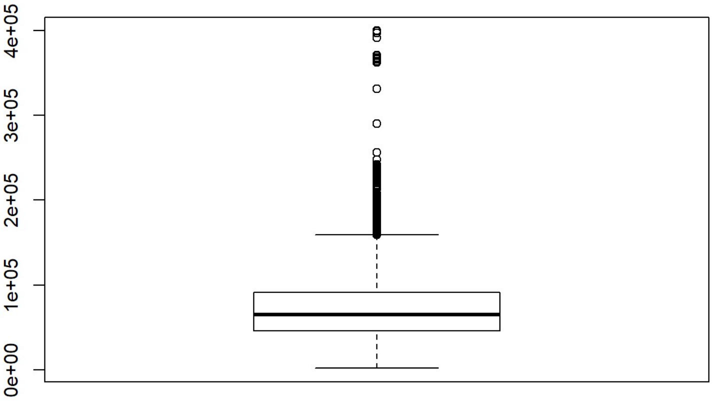
Section 1: Summarizing the data.

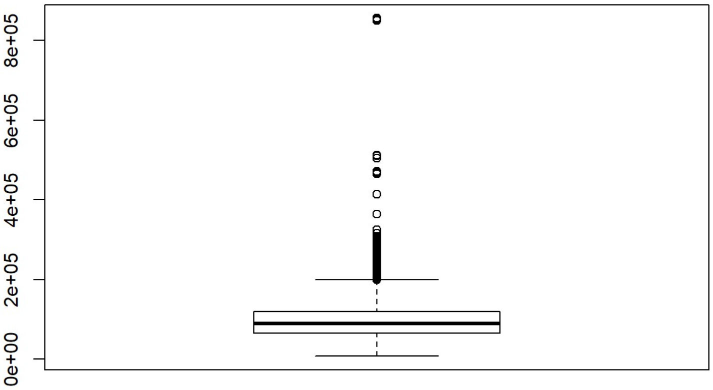
(1)Explore the data by summarizing it and graphing it.

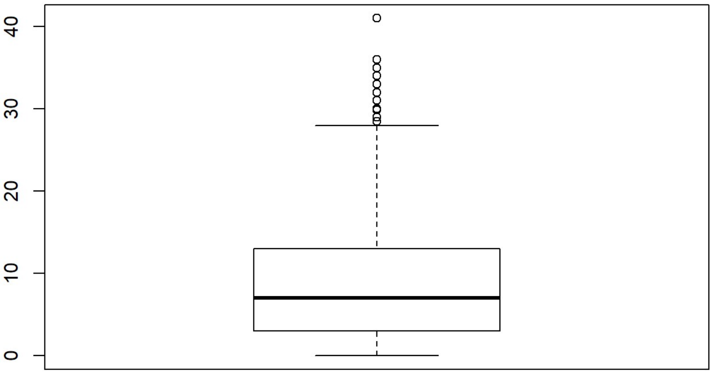
```
##   BAD LOAN MORTDUE  VALUE  REASON    JOB  YOJ DEROG DELINQ     CLAGE NINQ CLNO
## 1   1 1100   25860  39025 HomeImp  Other 10.5     0      0  94.36667    1    9
## 2   1 1300   70053  68400 HomeImp  Other  7.0     0      2 121.83333    0   14
## 3   1 1500   13500  16700 HomeImp  Other  4.0     0      0 149.46667    1   10
## 4   1 1500      NA     NA                 NA    NA     NA        NA   NA   NA
## 5   0 1700   97800 112000 HomeImp Office  3.0     0      0  93.33333    0   14
## 6   1 1700   30548  40320 HomeImp  Other  9.0     0      0 101.46600    1    8
##    DEBTINC
## 1       NA
## 2       NA
## 3       NA
## 4       NA
## 5       NA
## 6 37.11361
```

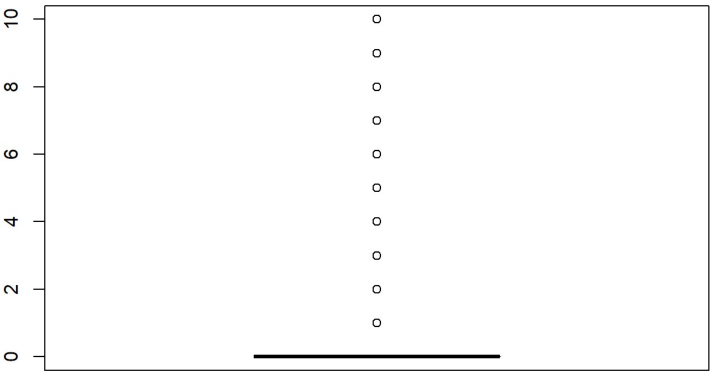
```
##      BAD  LOAN MORTDUE VALUE  REASON   JOB YOJ DEROG DELINQ    CLAGE NINQ CLNO
## 5955   0 88900   48919 93371 DebtCon Other  15     0      1 205.6502    0   15
## 5956   0 88900   57264 90185 DebtCon Other  16     0      0 221.8087    0   16
## 5957   0 89000   54576 92937 DebtCon Other  16     0      0 208.6921    0   15
## 5958   0 89200   54045 92924 DebtCon Other  15     0      0 212.2797    0   15
## 5959   0 89800   50370 91861 DebtCon Other  14     0      0 213.8927    0   16
## 5960   0 89900   48811 88934 DebtCon Other  15     0      0 219.6010    0   16
##       DEBTINC
## 5955 34.81826
## 5956 36.11235
## 5957 35.85997
## 5958 35.55659
## 5959 34.34088
## 5960 34.57152
```

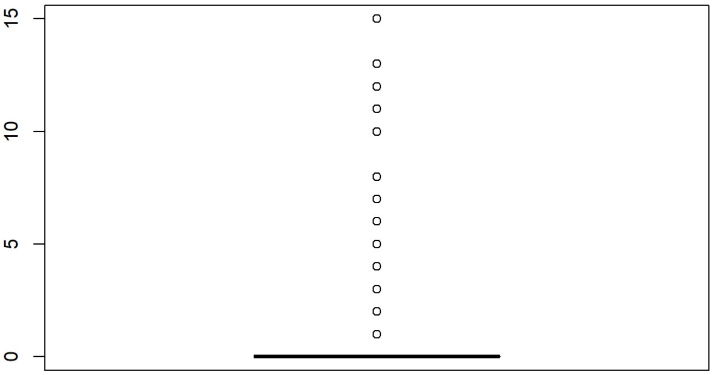
```
## 'data.frame':    5960 obs. of  13 variables:
##  $ BAD    : int  1 1 1 1 0 1 1 1 1 1 ...
##  $ LOAN   : int  1100 1300 1500 1500 1700 1700 1800 1800 2000 2000 ...
##  $ MORTDUE: num  25860 70053 13500 NA 97800 ...
##  $ VALUE  : num  39025 68400 16700 NA 112000 ...
##  $ REASON : Factor w/ 3 levels "","DebtCon","HomeImp": 3 3 3 1 3 3 3 3 3 3 ...
##  $ JOB    : Factor w/ 7 levels "","Mgr","Office",..: 4 4 4 1 3 4 4 4 4 6 ...
##  $ YOJ    : num  10.5 7 4 NA 3 9 5 11 3 16 ...
##  $ DEROG  : int  0 0 0 NA 0 0 3 0 0 0 ...
##  $ DELINQ : int  0 2 0 NA 0 0 2 0 2 0 ...
##  $ CLAGE  : num  94.4 121.8 149.5 NA 93.3 ...
##  $ NINQ   : int  1 0 1 NA 0 1 1 0 1 0 ...
##  $ CLNO   : int  9 14 10 NA 14 8 17 8 12 13 ...
##  $ DEBTINC: num  NA NA NA NA NA ...
##       BAD              LOAN          MORTDUE           VALUE
##  Min.   :0.0000   Min.   : 1100   Min.   :  2063   Min.   :  8000
##  1st Qu.:0.0000   1st Qu.:11100   1st Qu.: 46276   1st Qu.: 66076
##  Median :0.0000   Median :16300   Median : 65019   Median : 89236
##  Mean   :0.1995   Mean   :18608   Mean   : 73761   Mean   :101776
##  3rd Qu.:0.0000   3rd Qu.:23300   3rd Qu.: 91488   3rd Qu.:119824
##  Max.   :1.0000   Max.   :89900   Max.   :399550   Max.   :855909
##                                   NA's   :518      NA's   :112
##      REASON          JOB           YOJ             DEROG
##         : 252          : 279   Min.   : 0.000   Min.   : 0.0000
##  DebtCon:3928   Mgr    : 767   1st Qu.: 3.000   1st Qu.: 0.0000
##  HomeImp:1780   Office : 948   Median : 7.000   Median : 0.0000
##                 Other  :2388   Mean   : 8.922   Mean   : 0.2546
##                 ProfExe:1276   3rd Qu.:13.000   3rd Qu.: 0.0000
##                 Sales  : 109   Max.   :41.000   Max.   :10.0000
##                 Self   : 193   NA's   :515      NA's   :708
##      DELINQ            CLAGE             NINQ             CLNO
##  Min.   : 0.0000   Min.   :   0.0   Min.   : 0.000   Min.   : 0.0
##  1st Qu.: 0.0000   1st Qu.: 115.1   1st Qu.: 0.000   1st Qu.:15.0
##  Median : 0.0000   Median : 173.5   Median : 1.000   Median :20.0
##  Mean   : 0.4494   Mean   : 179.8   Mean   : 1.186   Mean   :21.3
##  3rd Qu.: 0.0000   3rd Qu.: 231.6   3rd Qu.: 2.000   3rd Qu.:26.0
##  Max.   :15.0000   Max.   :1168.2   Max.   :17.000   Max.   :71.0
##  NA's   :580       NA's   :308      NA's   :510      NA's   :222
##     DEBTINC
##  Min.   :  0.5245
##  1st Qu.: 29.1400
##  Median : 34.8183
##  Mean   : 33.7799
##  3rd Qu.: 39.0031
##  Max.   :203.3121
##  NA's   :1267
```

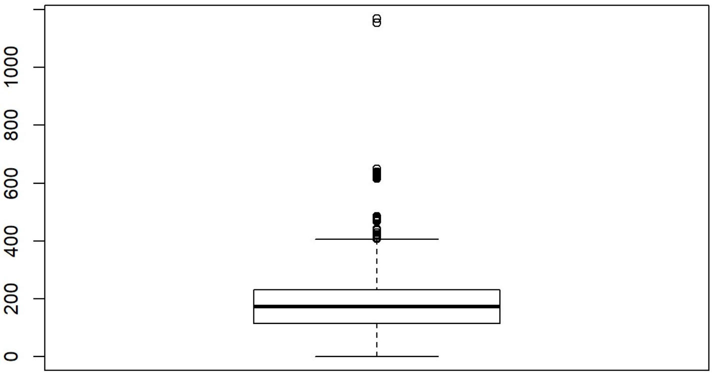

0e+00
1e+05
2e+05
3e+05
4e+05

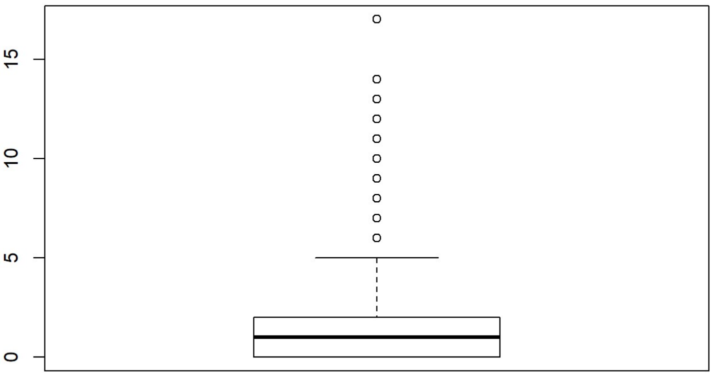

0e+00
2e+05
4e+05
6e+05
8e+05

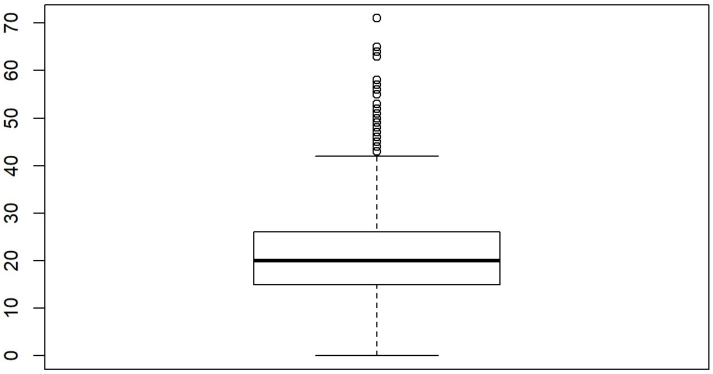

0
10
20
30
40

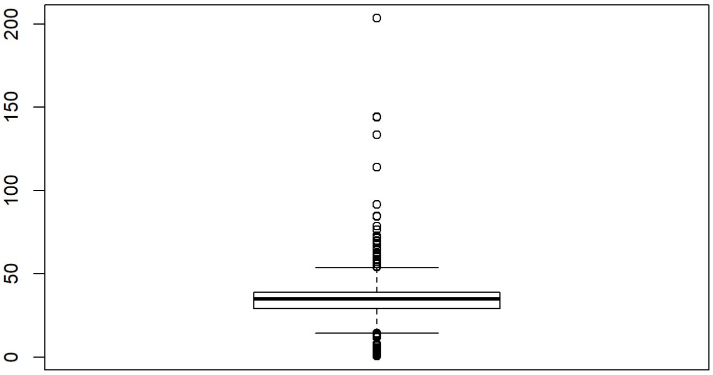

0
2
4
6
8
10


0
5
10
15


0
200
400
600
800
1000







(1)(b) Create a flag variable for each missing numeric variable. The variable name should begin with "M_".

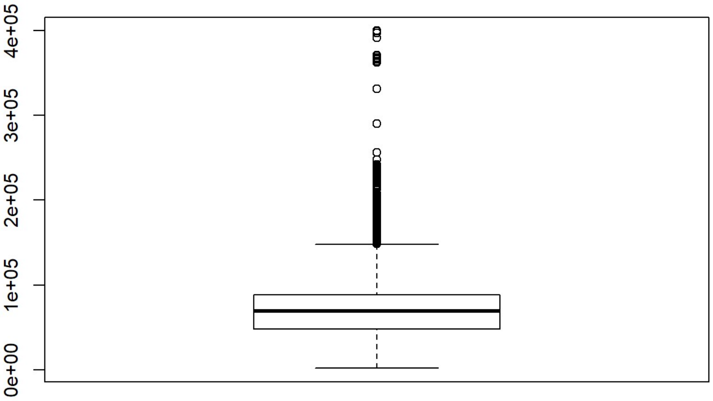
(1)(c) Create an imputed variable for each numeric variable that has a missing value. Fill in the missing value with the mean. The variable name should begin with “IMP_”

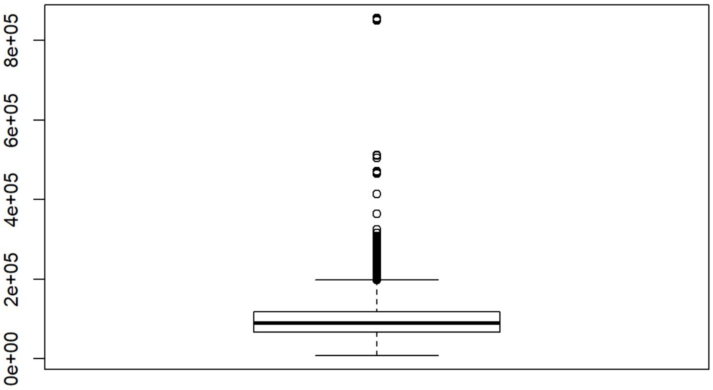
```
##   BAD LOAN MORTDUE  VALUE  REASON    JOB  YOJ DEROG DELINQ     CLAGE NINQ CLNO
## 1   1 1100   25860  39025 HomeImp  Other 10.5     0      0  94.36667    1    9
## 2   1 1300   70053  68400 HomeImp  Other  7.0     0      2 121.83333    0   14
## 3   1 1500   13500  16700 HomeImp  Other  4.0     0      0 149.46667    1   10
## 4   1 1500      NA     NA                  NA    NA     NA        NA   NA   NA
## 5   0 1700   97800 112000 HomeImp Office  3.0     0      0  93.33333    0   14
## 6   1 1700   30548  40320 HomeImp  Other  9.0     0      0 101.46600    1    8
##    DEBTINC M_MORTDUE M_VALUE M_YOJ M_DEROG M_DELINQ M_CLAGE M_NINQ M_CLNO
## 1       NA         0       0     0       0        0       0      0      0
## 2       NA         0       0     0       0        0       0      0      0
## 3       NA         0       0     0       0        0       0      0      0
## 4       NA         1       1     1       1        1       1      1      1
## 5       NA         0       0     0       0        0       0      0      0
## 6 37.11361         0       0     0       0        0       0      0      0
##   M_DEBTINC IMP_MORTDUE
## 1         1       25860
## 2         1       70053
## 3         1       13500
## 4         1       73761
## 5         1       97800
## 6         0       30548
```

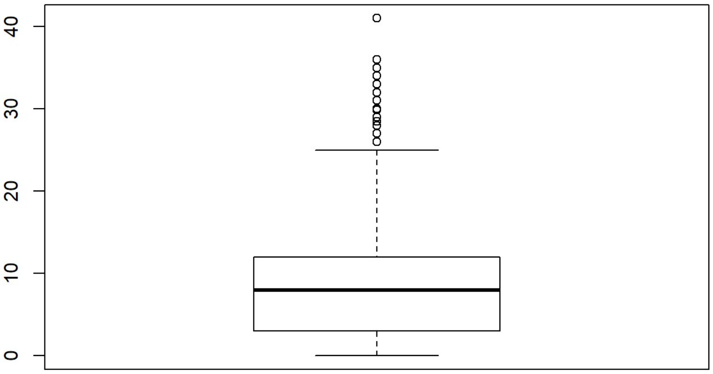
(1)(d) Create an imputed variable for each categorical variable that has a missing value. Fill in the missing value with the the value “UNKNOWN”. The variable name

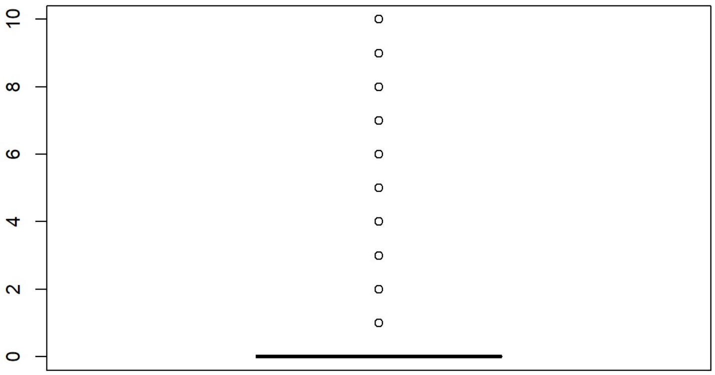
```
 ##   BAD LOAN MORTDUE  VALUE  REASON    JOB  YOJ DEROG DELINQ     CLAGE NINQ CLNO ## 1   1 1100   25860  39025 HomeImp  Other 10.5     0      0  9
```

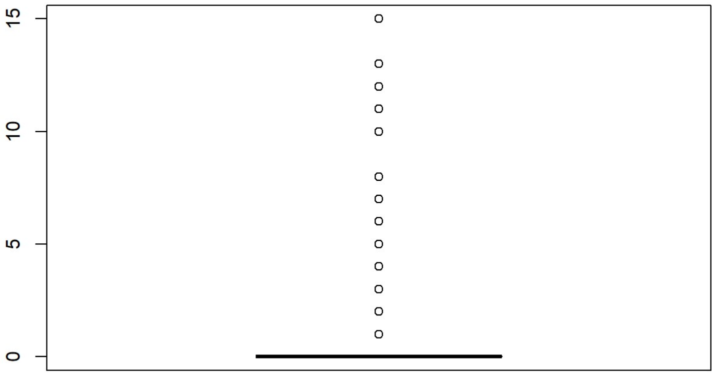
(1)(e) Identify any outliers and fix them in a method similar to those presented in the video. After variables are fixed, remove the original variables

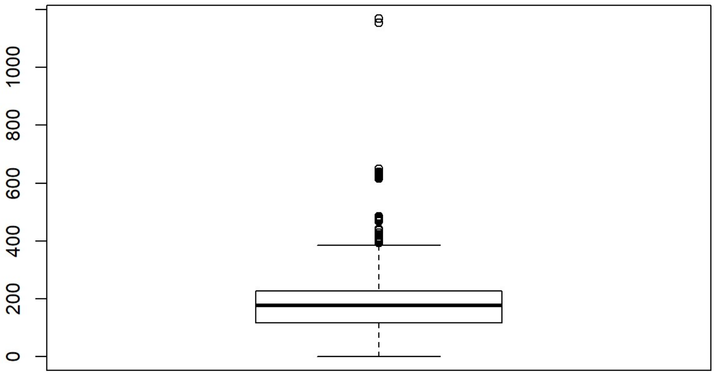

0e+00
1e+05
2e+05
3e+05
4e+05

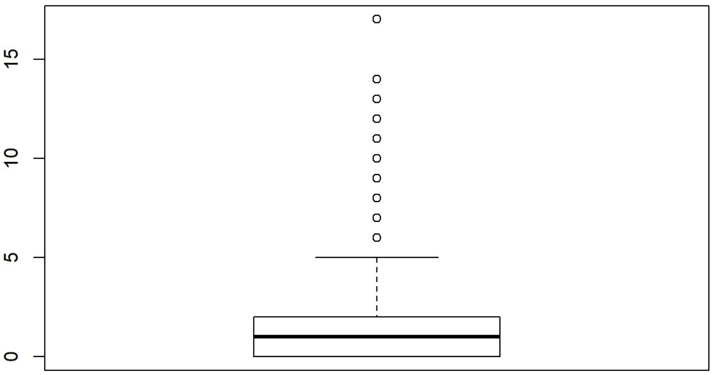

0e+00
2e+05
4e+05
6e+05
8e+05

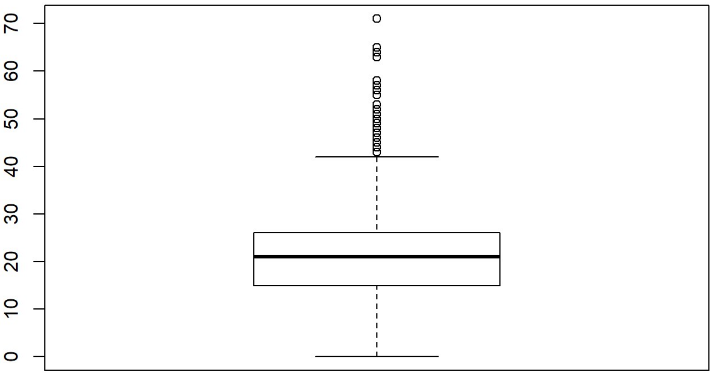

0
10
20
30
40

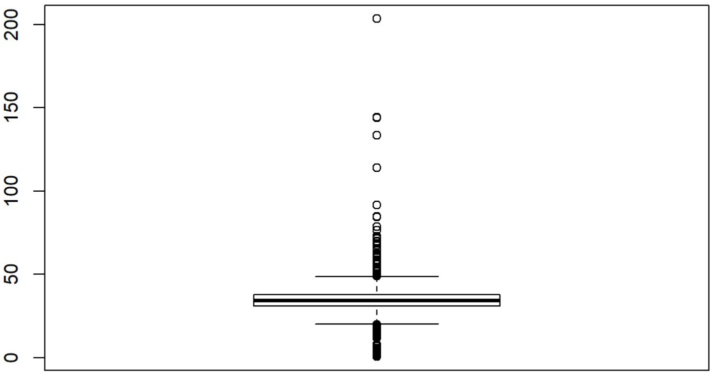

0
2
4
6
8
10


0
5
10
15


0
200
400
600
800
1000


0
5
10
15


0
10
20
30
40
50
60
70


0
50
100
150
200

```
##       BAD              LOAN          MORTDUE           VALUE       
##  Min.   :0.0000   Min.   : 1100   Min.   :  2063   Min.   :  8000  
##  1st Qu.:0.0000   1st Qu.:11100   1st Qu.: 46276   1st Qu.: 66076  
##  Median :0.0000   Median :16300   Median : 65019   Median : 89236  
##  Mean   :0.1995   Mean   :18608   Mean   : 73761   Mean   :101776  
##  3rd Qu.:0.0000   3rd Qu.:23300   3rd Qu.: 91488   3rd Qu.:119824  
##  Max.   :1.0000   Max.   :89900   Max.   :399550   Max.   :855909  
##                                   NA's   :518      NA's   :112     
##      REASON          JOB             YOJ             DEROG        
##         : 252          : 279   Min.   : 0.000   Min.   : 0.0000  
##  DebtCon:3928   Mgr    : 767   1st Qu.: 3.000   1st Qu.: 0.0000  
##  HomeImp:1780   Office : 948   Median : 7.000   Median : 0.0000  
##                 Other  :2388   Mean   : 8.922   Mean   : 0.2546  
##                 ProfExe:1276   3rd Qu.:13.000   3rd Qu.: 0.0000  
##                 Sales  : 109   Max.   :41.000   Max.   :10.0000  
##                 Self   : 193   NA's   :515      NA's   :708      
##      DELINQ            CLAGE             NINQ             CLNO     
##  Min.   : 0.0000   Min.   :   0.0   Min.   : 0.000   Min.   : 0.0  
##  1st Qu.: 0.0000   1st Qu.: 115.1   1st Qu.: 0.000   1st Qu.:15.0  
##  Median : 0.0000   Median : 173.5   Median : 1.000   Median :20.0  
##  Mean   : 0.4494   Mean   : 179.8   Mean   : 1.186   Mean   :21.3  
##  3rd Qu.: 0.0000   3rd Qu.: 231.6   3rd Qu.: 2.000   3rd Qu.:26.0  
##  Max.   :15.0000   Max.   :1168.2   Max.   :17.000   Max.   :71.0  
##  NA's   :580       NA's   :308      NA's   :510      NA's   :222   
##     DEBTINC           M_MORTDUE          M_VALUE            M_YOJ        
##  Min.   :  0.5245   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000  
##  1st Qu.: 29.1400   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000  
##  Median : 34.8183   Median :0.00000   Median :0.00000   Median :0.00000  
##  Mean   : 33.7799   Mean   :0.08691   Mean   :0.01879   Mean   :0.08641  
##  3rd Qu.: 39.0031   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000  
##  Max.   :203.3121   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000  
##  NA's   :1267                                                            
##     M_DEROG          M_DELINQ          M_CLAGE            M_NINQ       
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000  
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000  
##  Median :0.0000   Median :0.00000   Median :0.00000   Median :0.00000  
##  Mean   :0.1188   Mean   :0.09732   Mean   :0.05168   Mean   :0.08557  
##  3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000  
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000  
##                                                                        
##      M_CLNO          M_DEBTINC       IMP_MORTDUE       IMP_VALUE     
##  Min.   :0.00000   Min.   :0.0000   Min.   :  2063   Min.   :  8000  
##  1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.: 48139   1st Qu.: 66490  
##  Median :0.00000   Median :0.0000   Median : 69529   Median : 90000  
##  Mean   :0.03725   Mean   :0.2126   Mean   : 72971   Mean   :100700  
##  3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.: 88200   3rd Qu.:119005  
##  Max.   :1.00000   Max.   :1.0000   Max.   :201205   Max.   :272308  
##                                                                      
##     IMP_YOJ         IMP_DEROG        IMP_DELINQ       IMP_CLAGE    
##  Min.   : 0.000   Min.   :0.0000   Min.   :0.0000   Min.   :  0.0  
##  1st Qu.: 3.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:117.4  
##  Median : 8.000   Median :0.0000   Median :0.0000   Median :178.1  
##  Mean   : 8.913   Mean   :0.1838   Mean   :0.3613   Mean   :178.9  
##  3rd Qu.:12.000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:227.1  
##  Max.   :30.647   Max.   :2.6197   Max.   :3.6435   Max.   :430.5  
##                                                                    
##     IMP_NINQ        IMP_CLNO      IMP_DEBTINC     IMP_REASON       
##  Min.   :0.000   Min.   : 0.00   Min.   :10.93   Length:5960       
##  1st Qu.:0.000   1st Qu.:15.00   1st Qu.:30.76   Class :character  
##  Median :1.000   Median :21.00   Median :34.00   Mode  :character  
##  Mean   :1.118   Mean   :21.24   Mean   :33.74                     
##  3rd Qu.:2.000   3rd Qu.:26.00   3rd Qu.:37.95                     
##  Max.   :6.132   Max.   :51.13   Max.   :56.73                     
##                                                                    
##    IMP_JOB         
##  Length:5960       
##  Class :character  
##  Mode  :character  
##                    
##                    
##                    
## 
```

(1)(f) Remove the original fields

```
##       BAD              LOAN          M_MORTDUE          M_VALUE
##  Min.   :0.0000   Min.   : 1100   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.0000   1st Qu.:11100   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.0000   Median :16300   Median :0.00000   Median :0.00000
##  Mean   :0.1995   Mean   :18608   Mean   :0.08691   Mean   :0.01879
##  3rd Qu.:0.0000   3rd Qu.:23300   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.0000   Max.   :89900   Max.   :1.00000   Max.   :1.00000
##      M_YOJ            M_DEROG          M_DELINQ          M_CLAGE
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.00000   Median :0.0000   Median :0.00000   Median :0.00000
##  Mean   :0.08641   Mean   :0.1188   Mean   :0.09732   Mean   :0.05168
##  3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.00000   Max.   :1.00000
##      M_NINQ            M_CLNO          M_DEBTINC       IMP_MORTDUE
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.0000   Min.   :  2063
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.: 48139
##  Median :0.00000   Median :0.00000   Median :0.0000   Median : 69529
##  Mean   :0.08557   Mean   :0.03725   Mean   :0.2126   Mean   : 72971
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.: 88200
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.0000   Max.   :201205
##    IMP_VALUE         IMP_YOJ         IMP_DEROG        IMP_DELINQ
##  Min.   :  8000   Min.   : 0.000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.: 66490   1st Qu.: 3.000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median : 90000   Median : 8.000   Median :0.0000   Median :0.0000
##  Mean   :100700   Mean   : 8.913   Mean   :0.1838   Mean   :0.3613
##  3rd Qu.:119005   3rd Qu.:12.000   3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :272308   Max.   :30.647   Max.   :2.6197   Max.   :3.6435
##    IMP_CLAGE        IMP_NINQ        IMP_CLNO      IMP_DEBTINC
##  Min.   :  0.0   Min.   :0.000   Min.   : 0.00   Min.   :10.93
##  1st Qu.:117.4   1st Qu.:0.000   1st Qu.:15.00   1st Qu.:30.76
##  Median :178.1   Median :1.000   Median :21.00   Median :34.00
##  Mean   :178.9   Mean   :1.118   Mean   :21.24   Mean   :33.74
##  3rd Qu.:227.1   3rd Qu.:2.000   3rd Qu.:26.00   3rd Qu.:37.95
##  Max.   :430.5   Max.   :6.132   Max.   :51.13   Max.   :56.73
##   IMP_REASON          IMP_JOB
##  Length:5960        Length:5960
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```