Imputation

오태환

2020-05-27

1) Importing data

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.6.3
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## √ ggplot2 3.3.0     √ purrr   0.3.3
## √ tibble  2.1.3     √ dplyr   0.8.4
## √ tidyr   1.0.2     √ stringr 1.4.0
## √ readr   1.3.1     √ forcats 0.5.0
```

```
## Warning: package 'tidyr' was built under R version 3.6.3
```

```
## Warning: package 'purrr' was built under R version 3.6.3
```

```
## Warning: package 'dplyr' was built under R version 3.6.3
```

```
## Warning: package 'forcats' was built under R version 3.6.3
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(mice)
```

```
## Warning: package 'mice' was built under R version 3.6.3
```

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

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

```
library(impute)
data = read_csv("C:/Users/dhxog/Desktop/ESC4-1/Final_Project/FS_2y_before_Bankruptcy_train.csv"
)
```

```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   Attr1 = col_double(),
##   Attr2 = col_double(),
##   Attr3 = col_double(),
##   Attr5 = col_double(),
##   Attr6 = col_double(),
##   Attr7 = col_double(),
##   Attr9 = col_double(),
##   Attr10 = col_double(),
##   Attr11 = col_double(),
##   Attr14 = col_double(),
##   Attr15 = col_double(),
##   Attr18 = col_double(),
##   Attr22 = col_double(),
##   Attr25 = col_double(),
##   Attr29 = col_double(),
##   Attr35 = col_double(),
##   Attr36 = col_double(),
##   Attr38 = col_double(),
##   Attr48 = col_double(),
##   Attr51 = col_double()
##   # ... with 4 more columns
## )
```

```
## See spec(...) for full column specifications.
```

```
## Warning: 22 parsing failures.
##  row    col expected actual
file
## 1533 Attr5  a double      ? 'C:/Users/dhxog/Desktop/ESC4-1/Final_Project/FS_2y_before_Bankru
ptcy_train.csv'
## 2104 Attr5  a double      ? 'C:/Users/dhxog/Desktop/ESC4-1/Final_Project/FS_2y_before_Bankru
ptcy_train.csv'
## 2587 Attr15 a double      ? 'C:/Users/dhxog/Desktop/ESC4-1/Final_Project/FS_2y_before_Bankru
ptcy_train.csv'
## 2761 Attr5  a double      ? 'C:/Users/dhxog/Desktop/ESC4-1/Final_Project/FS_2y_before_Bankru
ptcy_train.csv'
## 2912 Attr5  a double      ? 'C:/Users/dhxog/Desktop/ESC4-1/Final_Project/FS_2y_before_Bankru
ptcy_train.csv'
## .... ...... ........ ......
.....................................................................
## See problems(...) for more details.
```

2) Change ? to NA

```
data[data == "?"] = NA

for(i in 1:ncol(data)){
  data[,i] = data[,i] %>% unlist() %>% as.numeric
}

summary(data)
```

```
##      Attr1               Attr2            Attr3              Attr4
##  Min.   :-12.458000   Min.   : 0.0000   Min.   :-15.48700   Min.   :   -0.045
##  1st Qu.:  0.001533   1st Qu.: 0.2636   1st Qu.:  0.01916   1st Qu.:    1.043
##  Median :  0.041770   Median : 0.4662   Median :  0.19727   Median :    1.585
##  Mean   :  0.045147   Mean   : 0.5423   Mean   :  0.18405   Mean   :    8.635
##  3rd Qu.:  0.112150   3rd Qu.: 0.6897   3rd Qu.:  0.40950   3rd Qu.:    2.868
##  Max.   :  9.803700   Max.   :40.1570   Max.   : 22.76900   Max.   :27146.000
##                                                             NA's   :28
##      Attr5               Attr6              Attr7
##  Min.   :-228990.0   Min.   :-117.4200   Min.   :-12.45800
##  1st Qu.:    -51.5   1st Qu.:  -0.0004   1st Qu.:  0.00296
##  Median :      0.0   Median :   0.0000   Median :  0.04911
##  Mean   :    146.8   Mean   :   0.0156   Mean   :  0.06249
##  3rd Qu.:     55.4   3rd Qu.:   0.0622   3rd Qu.:  0.12759
##  Max.   :1034100.0   Max.   : 322.2000   Max.   : 38.61800
##  NA's   :15
##      Attr8               Attr9              Attr10             Attr11
##  Min.   :   -0.975   Min.   :  -0.0324   Min.   :-39.1560   Min.   :-12.24400
##  1st Qu.:    0.429   1st Qu.:   1.0076   1st Qu.:  0.2957   1st Qu.:  0.00936
##  Median :    1.093   Median :   1.1621   Median :  0.5114   Median :  0.06294
##  Mean   :   13.091   Mean   :   1.9735   Mean   :  0.4442   Mean   :  0.07746
##  3rd Qu.:    2.684   3rd Qu.:   1.9620   3rd Qu.:  0.7142   3rd Qu.:  0.14283
##  Max.   :27145.000   Max.   :1704.8000   Max.   : 12.6020   Max.   : 38.61800
##  NA's   :15
##      Attr12              Attr13              Attr14             Attr15
##  Min.   :-1236.400   Min.   :-1460.6000   Min.   :-12.45800   Min.   :-789840
##  1st Qu.:    0.008   1st Qu.:    0.0214   1st Qu.:  0.00299   1st Qu.:    222
##  Median :    0.145   Median :    0.0647   Median :  0.04916   Median :    913
##  Mean   :    1.030   Mean   :    0.2195   Mean   :  0.06251   Mean   :   3853
##  3rd Qu.:    0.522   3rd Qu.:    0.1302   3rd Qu.:  0.12769   3rd Qu.:   2408
##  Max.   : 3340.900   Max.   : 2707.7000   Max.   : 38.61800   Max.   :8085500
##  NA's   :28          NA's   :14                               NA's   :5
##      Attr16             Attr17             Attr18
##  Min.   :-134.890   Min.   :    0.001   Min.   :-12.45800
##  1st Qu.:   0.062   1st Qu.:    1.447   1st Qu.:  0.00299
##  Median :   0.223   Median :    2.143   Median :  0.04916
##  Mean   :   1.502   Mean   :   14.184   Mean   :  0.06982
##  3rd Qu.:   0.608   3rd Qu.:    3.773   3rd Qu.:  0.12769
##  Max.   :4401.300   Max.   :27146.000   Max.   : 50.26600
##  NA's   :15         NA's   :15
##      Attr19              Attr20             Attr21             Attr22
##  Min.   :-1578.7000   Min.   :    0.00   Min.   : -1.1463   Min.   :-12.24400
##  1st Qu.:    0.0019   1st Qu.:   15.20   1st Qu.:  0.9225   1st Qu.:  0.00000
##  Median :    0.0314   Median :   35.39   Median :  1.0472   Median :  0.04958
##  Mean   :   -0.3497   Mean   :   64.08   Mean   :  1.2317   Mean   :  0.06998
##  3rd Qu.:    0.0842   3rd Qu.:   65.05   3rd Qu.:  1.2091   3rd Qu.:  0.12786
##  Max.   :  497.0200   Max.   :26606.00   Max.   :396.1600   Max.   : 38.61800
##  NA's   :14           NA's   :14         NA's   :112
##      Attr23              Attr24             Attr25             Attr26
##  Min.   :-1578.7000   Min.   :-289.1200   Min.   :-47.5310   Min.   :-134.890
##  1st Qu.:    0.0010   1st Qu.:   0.0053   1st Qu.:  0.1342   1st Qu.:   0.058
##  Median :    0.0261   Median :   0.1475   Median :  0.3908   Median :   0.204
##  Mean   :   -0.3592   Mean   :   0.3456   Mean   :  0.2762   Mean   :   1.258
##  3rd Qu.:    0.0729   3rd Qu.:   0.3621   3rd Qu.:  0.6159   3rd Qu.:   0.548
##  Max.   :  497.0200   Max.   : 400.5900   Max.   : 12.6020   Max.   :3594.600
##  NA's   :14           NA's   :149                            NA's   :15
##      Attr27              Attr28             Attr29             Attr30
```

```
##  Min.   :-109340.0   Min.   : -990.020   Min.   :-0.3565   Min.   :-1055.900
##  1st Qu.:      0.0   1st Qu.:    0.034   1st Qu.: 3.4125   1st Qu.:    0.085
##  Median :      1.0   Median :    0.463   Median : 3.9820   Median :    0.229
##  Mean   :   1312.0   Mean   :    6.859   Mean   : 3.9553   Mean   :    8.706
##  3rd Qu.:      5.4   3rd Qu.:    1.512   3rd Qu.: 4.5039   3rd Qu.:    0.429
##  Max.   :2037300.0   Max.   :11864.000   Max.   : 7.6009   Max.   :29526.000
##  NA's   :462         NA's   :162                           NA's   :14
##      Attr31              Attr32             Attr33            Attr34
##  Min.   :-1495.6000   Min.   :     0.00   Min.   :   0.000   Min.   :-756.500
##  1st Qu.:    0.0043   1st Qu.:    47.54   1st Qu.:   2.718   1st Qu.:   0.283
##  Median :    0.0383   Median :    81.18   Median :   4.451   Median :   1.958
##  Mean   :   -0.2740   Mean   :   260.77   Mean   :   8.855   Mean   :   5.251
##  3rd Qu.:    0.0954   3rd Qu.:   132.40   3rd Qu.:   7.550   3rd Qu.:   4.468
##  Max.   :  798.3200   Max.   :141510.00   Max.   :5534.100   Max.   :4260.200
##  NA's   :14           NA's   :72          NA's   :28         NA's   :15
##      Attr35             Attr36             Attr37              Attr38
##  Min.   :-4.79220   Min.   :   0.000   Min.   :   -3.715   Min.   :-39.1560
##  1st Qu.: 0.00104   1st Qu.:   1.040   1st Qu.:    1.089   1st Qu.:  0.4258
##  Median : 0.04669   Median :   1.548   Median :    3.088   Median :  0.6139
##  Mean   : 0.06238   Mean   :   2.153   Mean   :   70.807   Mean   :  0.5414
##  3rd Qu.: 0.12688   3rd Qu.:   2.269   3rd Qu.:   12.204   3rd Qu.:  0.7771
##  Max.   :38.61800   Max.   :1704.800   Max.   :24487.000   Max.   : 12.6020
##                                        NA's   :3100
##      Attr39              Attr40             Attr41
##  Min.   :-7522.000   Min.   :  -6.8769   Min.   :-1086.800
##  1st Qu.:    0.001   1st Qu.:   0.0506   1st Qu.:    0.025
##  Median :    0.030   Median :   0.1772   Median :    0.089
##  Mean   :   -1.461   Mean   :   2.3760   Mean   :    1.056
##  3rd Qu.:    0.083   3rd Qu.:   0.6579   3rd Qu.:    0.216
##  Max.   :  112.020   Max.   :2028.5000   Max.   : 3443.400
##  NA's   :14          NA's   :28          NA's   :142
##      Attr42              Attr43             Attr44
##  Min.   :-719.8000   Min.   :-115870.0   Min.   :-115870.0
##  1st Qu.:   0.0000   1st Qu.:     68.8   1st Qu.:     36.7
##  Median :   0.0324   Median :    103.6   Median :     58.1
##  Mean   :  -0.4004   Mean   :    893.8   Mean   :    829.7
##  3rd Qu.:   0.0837   3rd Qu.:    147.5   3rd Qu.:     85.8
##  Max.   : 160.1100   Max.   :3020000.0   Max.   :3020000.0
##  NA's   :14          NA's   :14          NA's   :14
##      Attr45              Attr46             Attr47
##  Min.   :-2834.900   Min.   :   -6.639   Min.   :    -3.63
##  1st Qu.:    0.011   1st Qu.:    0.617   1st Qu.:    15.92
##  Median :    0.235   Median :    1.043   Median :    38.15
##  Mean   :    4.143   Mean   :    7.797   Mean   :   132.47
##  3rd Qu.:    0.834   3rd Qu.:    1.961   3rd Qu.:    69.96
##  Max.   :10337.000   Max.   :27146.000   Max.   :140990.00
##  NA's   :418         NA's   :28          NA's   :57
##      Attr48              Attr49             Attr50              Attr51
##  Min.   :-13.81500   Min.   :-837.8600   Min.   :   -0.045   Min.   : 0.0000
##  1st Qu.: -0.04600   1st Qu.:  -0.0340   1st Qu.:    0.768   1st Qu.: 0.1855
##  Median :  0.00717   Median :   0.0044   Median :    1.227   Median : 0.3362
##  Mean   : -0.00027   Mean   :  -0.5807   Mean   :    7.458   Mean   : 0.4146
##  3rd Qu.:  0.08490   3rd Qu.:   0.0519   3rd Qu.:    2.229   3rd Qu.: 0.5279
##  Max.   : 33.53500   Max.   : 107.6800   Max.   :27146.000   Max.   :16.4870
##                      NA's   :14          NA's   :15
##      Attr52             Attr53              Attr54              Attr55
##  Min.   :  0.0000   Min.   :-1033.700   Min.   :-1033.700   Min.   :-504580
##  1st Qu.:  0.1299   1st Qu.:    0.685   1st Qu.:    0.949   1st Qu.:     19
```

```
##  Median :  0.2224   Median :    1.208   Median :    1.375   Median :    966
##  Mean   :  0.7528   Mean   :    5.948   Mean   :    7.607   Mean   :   8406
##  3rd Qu.:  0.3620   3rd Qu.:    2.241   3rd Qu.:    2.381   3rd Qu.:   4823
##  Max.   :387.7100   Max.   : 4784.100   Max.   :11678.000   Max.   :6123700
##  NA's   :60         NA's   :162         NA's   :162
##      Attr56              Attr57              Attr58             Attr59
##  Min.   :-7522.100   Min.   :-481.3100   Min.   :-30.8920   Min.   :-284.3800
##  1st Qu.:    0.004   1st Qu.:   0.0091   1st Qu.:  0.8821   1st Qu.:   0.0000
##  Median :    0.045   Median :   0.0980   Median :  0.9562   Median :   0.0026
##  Mean   :   -1.411   Mean   :   0.0769   Mean   :  1.1391   Mean   :   0.7366
##  3rd Qu.:    0.120   3rd Qu.:   0.2429   3rd Qu.:  0.9956   3rd Qu.:   0.2104
##  Max.   :  112.020   Max.   : 226.7600   Max.   :668.7500   Max.   :1661.0000
##  NA's   :14          NA's   :1           NA's   :10         NA's   :1
##      Attr60              Attr61              Attr62             Attr63
##  Min.   :     0.00   Min.   :     0.00   Min.   :  -14965   Min.   :  -0.024
##  1st Qu.:     5.40   1st Qu.:     4.24   1st Qu.:      44   1st Qu.:   2.966
##  Median :     9.56   Median :     6.25   Median :      75   Median :   4.847
##  Mean   :   134.16   Mean   :    31.35   Mean   :    2697   Mean   :   8.979
##  3rd Qu.:    19.61   3rd Qu.:     9.83   3rd Qu.:     122   3rd Qu.:   8.304
##  Max.   :251570.00   Max.   :108000.00   Max.   :10779000   Max.   :5662.400
##  NA's   :420         NA's   :20          NA's   :14         NA's   :28
##      Attr64             class
##  Min.   :    0.000   Min.   :0.00000
##  1st Qu.:    2.010   1st Qu.:0.00000
##  Median :    4.012   Median :0.00000
##  Mean   :   36.199   Mean   :0.05266
##  3rd Qu.:    9.086   3rd Qu.:0.00000
##  Max.   :21153.000   Max.   :1.00000
##  NA's   :162
```

3) NA가 많은 Attr37 제거 후 imputation 진행

3-1) MICE 패키지의 PMM방식 사용

```
tempdata = data %>% select(-Attr37)

imp = mice(tempdata, seed = 1234)
```

```
##
##  iter imp variable
##   1   1  Attr4  Attr5  Attr8  Attr12  Attr13  Attr15  Attr16  Attr19  Attr20  Attr21  Attr23
Attr24  Attr27  Attr28  Attr30  Attr31  Attr32  Attr33  Attr34  Attr39  Attr40  Attr41  Attr42
Attr43  Attr45  Attr47  Attr49  Attr50  Attr52  Attr53  Attr57  Attr58  Attr59  Attr60  Attr61
Attr62  Attr63  Attr64
##   1   2  Attr4  Attr5  Attr8  Attr12  Attr13  Attr15  Attr16  Attr19  Attr20  Attr21  Attr23
Attr24  Attr27  Attr28  Attr30  Attr31  Attr32  Attr33  Attr34  Attr39  Attr40  Attr41  Attr42
Attr43  Attr45  Attr47  Attr49  Attr50  Attr52  Attr53  Attr57  Attr58  Attr59  Attr60  Attr61
Attr62  Attr63  Attr64
##   1   3  Attr4  Attr5  Attr8  Attr12  Attr13  Attr15  Attr16  Attr19  Attr20  Attr21  Attr23
Attr24  Attr27  Attr28  Attr30  Attr31  Attr32  Attr33  Attr34  Attr39  Attr40  Attr41  Attr42
Attr43  Attr45  Attr47  Attr49  Attr50  Attr52  Attr53  Attr57  Attr58  Attr59  Attr60  Attr61
Attr62  Attr63  Attr64
##   1   4  Attr4  Attr5  Attr8  Attr12  Attr13  Attr15  Attr16  Attr19  Attr20  Attr21  Attr23
Attr24  Attr27  Attr28  Attr30  Attr31  Attr32  Attr33  Attr34  Attr39  Attr40  Attr41  Attr42
Attr43  Attr45  Attr47  Attr49  Attr50  Attr52  Attr53  Attr57  Attr58  Attr59  Attr60  Attr61
Attr62  Attr63  Attr64
##   1   5  Attr4  Attr5  Attr8  Attr12  Attr13  Attr15  Attr16  Attr19  Attr20  Attr21  Attr23
Attr24  Attr27  Attr28  Attr30  Attr31  Attr32  Attr33  Attr34  Attr39  Attr40  Attr41  Attr42
Attr43  Attr45  Attr47  Attr49  Attr50  Attr52  Attr53  Attr57  Attr58  Attr59  Attr60  Attr61
Attr62  Attr63  Attr64
##   2   1  Attr4  Attr5  Attr8  Attr12  Attr13  Attr15  Attr16  Attr19  Attr20  Attr21  Attr23
Attr24  Attr27  Attr28  Attr30  Attr31  Attr32  Attr33  Attr34  Attr39  Attr40  Attr41  Attr42
Attr43  Attr45  Attr47  Attr49  Attr50  Attr52  Attr53  Attr57  Attr58  Attr59  Attr60  Attr61
Attr62  Attr63  Attr64
##   2   2  Attr4  Attr5  Attr8  Attr12  Attr13  Attr15  Attr16  Attr19  Attr20  Attr21  Attr23
Attr24  Attr27  Attr28  Attr30  Attr31  Attr32  Attr33  Attr34  Attr39  Attr40  Attr41  Attr42
Attr43  Attr45  Attr47  Attr49  Attr50  Attr52  Attr53  Attr57  Attr58  Attr59  Attr60  Attr61
Attr62  Attr63  Attr64
##   2   3  Attr4  Attr5  Attr8  Attr12  Attr13  Attr15  Attr16  Attr19  Attr20  Attr21  Attr23
Attr24  Attr27  Attr28  Attr30  Attr31  Attr32  Attr33  Attr34  Attr39  Attr40  Attr41  Attr42
Attr43  Attr45  Attr47  Attr49  Attr50  Attr52  Attr53  Attr57  Attr58  Attr59  Attr60  Attr61
Attr62  Attr63  Attr64
##   2   4  Attr4  Attr5  Attr8  Attr12  Attr13  Attr15  Attr16  Attr19  Attr20  Attr21  Attr23
Attr24  Attr27  Attr28  Attr30  Attr31  Attr32  Attr33  Attr34  Attr39  Attr40  Attr41  Attr42
Attr43  Attr45  Attr47  Attr49  Attr50  Attr52  Attr53  Attr57  Attr58  Attr59  Attr60  Attr61
Attr62  Attr63  Attr64
##   2   5  Attr4  Attr5  Attr8  Attr12  Attr13  Attr15  Attr16  Attr19  Attr20  Attr21  Attr23
Attr24  Attr27  Attr28  Attr30  Attr31  Attr32  Attr33  Attr34  Attr39  Attr40  Attr41  Attr42
Attr43  Attr45  Attr47  Attr49  Attr50  Attr52  Attr53  Attr57  Attr58  Attr59  Attr60  Attr61
Attr62  Attr63  Attr64
##   3   1  Attr4  Attr5  Attr8  Attr12  Attr13  Attr15  Attr16  Attr19  Attr20  Attr21  Attr23
Attr24  Attr27  Attr28  Attr30  Attr31  Attr32  Attr33  Attr34  Attr39  Attr40  Attr41  Attr42
Attr43  Attr45  Attr47  Attr49  Attr50  Attr52  Attr53  Attr57  Attr58  Attr59  Attr60  Attr61
Attr62  Attr63  Attr64
##   3   2  Attr4  Attr5  Attr8  Attr12  Attr13  Attr15  Attr16  Attr19  Attr20  Attr21  Attr23
Attr24  Attr27  Attr28  Attr30  Attr31  Attr32  Attr33  Attr34  Attr39  Attr40  Attr41  Attr42
Attr43  Attr45  Attr47  Attr49  Attr50  Attr52  Attr53  Attr57  Attr58  Attr59  Attr60  Attr61
Attr62  Attr63  Attr64
##   3   3  Attr4  Attr5  Attr8  Attr12  Attr13  Attr15  Attr16  Attr19  Attr20  Attr21  Attr23
Attr24  Attr27  Attr28  Attr30  Attr31  Attr32  Attr33  Attr34  Attr39  Attr40  Attr41  Attr42
Attr43  Attr45  Attr47  Attr49  Attr50  Attr52  Attr53  Attr57  Attr58  Attr59  Attr60  Attr61
Attr62  Attr63  Attr64
##   3   4  Attr4  Attr5  Attr8  Attr12  Attr13  Attr15  Attr16  Attr19  Attr20  Attr21  Attr23
Attr24  Attr27  Attr28  Attr30  Attr31  Attr32  Attr33  Attr34  Attr39  Attr40  Attr41  Attr42
Attr43  Attr45  Attr47  Attr49  Attr50  Attr52  Attr53  Attr57  Attr58  Attr59  Attr60  Attr61
```

```
Attr62  Attr63  Attr64
##   3   5  Attr4  Attr5  Attr8  Attr12  Attr13  Attr15  Attr16  Attr19  Attr20  Attr21  Attr23
Attr24  Attr27  Attr28  Attr30  Attr31  Attr32  Attr33  Attr34  Attr39  Attr40  Attr41  Attr42
Attr43  Attr45  Attr47  Attr49  Attr50  Attr52  Attr53  Attr57  Attr58  Attr59  Attr60  Attr61
Attr62  Attr63  Attr64
##   4   1  Attr4  Attr5  Attr8  Attr12  Attr13  Attr15  Attr16  Attr19  Attr20  Attr21  Attr23
Attr24  Attr27  Attr28  Attr30  Attr31  Attr32  Attr33  Attr34  Attr39  Attr40  Attr41  Attr42
Attr43  Attr45  Attr47  Attr49  Attr50  Attr52  Attr53  Attr57  Attr58  Attr59  Attr60  Attr61
Attr62  Attr63  Attr64
##   4   2  Attr4  Attr5  Attr8  Attr12  Attr13  Attr15  Attr16  Attr19  Attr20  Attr21  Attr23
Attr24  Attr27  Attr28  Attr30  Attr31  Attr32  Attr33  Attr34  Attr39  Attr40  Attr41  Attr42
Attr43  Attr45  Attr47  Attr49  Attr50  Attr52  Attr53  Attr57  Attr58  Attr59  Attr60  Attr61
Attr62  Attr63  Attr64
##   4   3  Attr4  Attr5  Attr8  Attr12  Attr13  Attr15  Attr16  Attr19  Attr20  Attr21  Attr23
Attr24  Attr27  Attr28  Attr30  Attr31  Attr32  Attr33  Attr34  Attr39  Attr40  Attr41  Attr42
Attr43  Attr45  Attr47  Attr49  Attr50  Attr52  Attr53  Attr57  Attr58  Attr59  Attr60  Attr61
Attr62  Attr63  Attr64
##   4   4  Attr4  Attr5  Attr8  Attr12  Attr13  Attr15  Attr16  Attr19  Attr20  Attr21  Attr23
Attr24  Attr27  Attr28  Attr30  Attr31  Attr32  Attr33  Attr34  Attr39  Attr40  Attr41  Attr42
Attr43  Attr45  Attr47  Attr49  Attr50  Attr52  Attr53  Attr57  Attr58  Attr59  Attr60  Attr61
Attr62  Attr63  Attr64
##   4   5  Attr4  Attr5  Attr8  Attr12  Attr13  Attr15  Attr16  Attr19  Attr20  Attr21  Attr23
Attr24  Attr27  Attr28  Attr30  Attr31  Attr32  Attr33  Attr34  Attr39  Attr40  Attr41  Attr42
Attr43  Attr45  Attr47  Attr49  Attr50  Attr52  Attr53  Attr57  Attr58  Attr59  Attr60  Attr61
Attr62  Attr63  Attr64
##   5   1  Attr4  Attr5  Attr8  Attr12  Attr13  Attr15  Attr16  Attr19  Attr20  Attr21  Attr23
Attr24  Attr27  Attr28  Attr30  Attr31  Attr32  Attr33  Attr34  Attr39  Attr40  Attr41  Attr42
Attr43  Attr45  Attr47  Attr49  Attr50  Attr52  Attr53  Attr57  Attr58  Attr59  Attr60  Attr61
Attr62  Attr63  Attr64
##   5   2  Attr4  Attr5  Attr8  Attr12  Attr13  Attr15  Attr16  Attr19  Attr20  Attr21  Attr23
Attr24  Attr27  Attr28  Attr30  Attr31  Attr32  Attr33  Attr34  Attr39  Attr40  Attr41  Attr42
Attr43  Attr45  Attr47  Attr49  Attr50  Attr52  Attr53  Attr57  Attr58  Attr59  Attr60  Attr61
Attr62  Attr63  Attr64
##   5   3  Attr4  Attr5  Attr8  Attr12  Attr13  Attr15  Attr16  Attr19  Attr20  Attr21  Attr23
Attr24  Attr27  Attr28  Attr30  Attr31  Attr32  Attr33  Attr34  Attr39  Attr40  Attr41  Attr42
Attr43  Attr45  Attr47  Attr49  Attr50  Attr52  Attr53  Attr57  Attr58  Attr59  Attr60  Attr61
Attr62  Attr63  Attr64
##   5   4  Attr4  Attr5  Attr8  Attr12  Attr13  Attr15  Attr16  Attr19  Attr20  Attr21  Attr23
Attr24  Attr27  Attr28  Attr30  Attr31  Attr32  Attr33  Attr34  Attr39  Attr40  Attr41  Attr42
Attr43  Attr45  Attr47  Attr49  Attr50  Attr52  Attr53  Attr57  Attr58  Attr59  Attr60  Attr61
Attr62  Attr63  Attr64
##   5   5  Attr4  Attr5  Attr8  Attr12  Attr13  Attr15  Attr16  Attr19  Attr20  Attr21  Attr23
Attr24  Attr27  Attr28  Attr30  Attr31  Attr32  Attr33  Attr34  Attr39  Attr40  Attr41  Attr42
Attr43  Attr45  Attr47  Attr49  Attr50  Attr52  Attr53  Attr57  Attr58  Attr59  Attr60  Attr61
Attr62  Attr63  Attr64
```

```
## Warning: Number of logged events: 957
```

아직 안지워지는 NA가 있다! 이것은 상당히 높은 Correlation을 가진 Column들이 있기 때문이다. 어떤 것들이 Correlation이 높은지 찾아보자

```
cortable = cor(data, use = "pairwise.complete.obs")


for(i in 1:ncol(cortable)){
  temp = c()
  for(j in 1:nrow(cortable)){
    if(cortable[j,i] > 0.99 & j > i ){
      temp = cbind(temp, paste0("Attr", j))
    }
  }
  if(length(temp) >0){
  cat("High Relative Comp for Attr", i, ":", temp, '₩n')
  }
}
```

```
## High Relative Comp for Attr 4 : Attr46 Attr50
## High Relative Comp for Attr 7 : Attr11 Attr14
## High Relative Comp for Attr 8 : Attr17
## High Relative Comp for Attr 11 : Attr14
## High Relative Comp for Attr 16 : Attr26
## High Relative Comp for Attr 19 : Attr23
## High Relative Comp for Attr 28 : Attr54
## High Relative Comp for Attr 32 : Attr52
## High Relative Comp for Attr 39 : Attr56
## High Relative Comp for Attr 43 : Attr44
## High Relative Comp for Attr 46 : Attr50
```

```
complete.data = complete(imp)

for(i in 1:ncol(complete.data)){
  if(sum(is.na(complete.data[,i] == T))){
    cat("Attr", i, "has", sum(is.na(complete.data[,i])), "NAs", "₩n")
  }
}
```

```
## Attr 17 has 15 NAs
## Attr 26 has 15 NAs
## Attr 43 has 14 NAs
## Attr 45 has 28 NAs
## Attr 53 has 162 NAs
## Attr 55 has 14 NAs
```

```
data_noimp = read_csv("C:/Users/dhxog/Desktop/ESC4-1/Final_Project/data_removed_without_imputat
ion.csv")
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   X1 = col_double(),
##   Attr1 = col_double(),
##   Attr5 = col_double(),
##   Attr6 = col_double(),
##   Attr9 = col_double(),
##   Attr10 = col_double(),
##   Attr15 = col_double(),
##   Attr18 = col_double(),
##   Attr29 = col_double(),
##   Attr55 = col_double(),
##   Attr57 = col_double(),
##   Attr59 = col_double(),
##   class = col_double()
## )
```

```
## See spec(...) for full column specifications.
```

```
## Warning: 22 parsing failures.
##  row    col expected actual
file
## 1533 Attr5  a double      ? 'C:/Users/dhxog/Desktop/ESC4-1/Final_Project/data_removed_withou
t_imputation.csv'
## 2104 Attr5  a double      ? 'C:/Users/dhxog/Desktop/ESC4-1/Final_Project/data_removed_withou
t_imputation.csv'
## 2587 Attr15 a double      ? 'C:/Users/dhxog/Desktop/ESC4-1/Final_Project/data_removed_withou
t_imputation.csv'
## 2761 Attr5  a double      ? 'C:/Users/dhxog/Desktop/ESC4-1/Final_Project/data_removed_withou
t_imputation.csv'
## 2912 Attr5  a double      ? 'C:/Users/dhxog/Desktop/ESC4-1/Final_Project/data_removed_withou
t_imputation.csv'
## .... ...... ........ ......
.....................................................................
## See problems(...) for more details.
```

```
colnames(data_noimp)
```

```
##  [1] "X1"     "Attr1"  "Attr5"  "Attr6"  "Attr9"  "Attr10" "Attr15" "Attr17"
##  [9] "Attr18" "Attr19" "Attr20" "Attr21" "Attr26" "Attr27" "Attr29" "Attr37"
## [17] "Attr41" "Attr42" "Attr45" "Attr46" "Attr47" "Attr54" "Attr55" "Attr57"
## [25] "Attr59" "Attr60" "Attr61" "Attr63" "Attr64" "class"
```

```
tempdata = tempdata %>% select(-c("Attr4", "Attr50", "Attr11", "Attr14", "Attr8", "Attr16", "At
tr23", "Attr28",
                          "Attr52","Attr56", "Attr44"))

imp = mice(tempdata, seed = 1234)
```

```
##
##  iter imp variable
##   1   1  Attr5  Attr12  Attr13  Attr15  Attr17  Attr19  Attr20  Attr21  Attr24  Attr26  Attr
27  Attr30  Attr31  Attr32  Attr33  Attr34  Attr39  Attr40  Attr41  Attr42  Attr43  Attr45  Att
r46  Attr47  Attr49  Attr53  Attr54  Attr57  Attr58  Attr59  Attr60  Attr61  Attr62  Attr63  At
tr64
##   1   2  Attr5  Attr12  Attr13  Attr15  Attr17  Attr19  Attr20  Attr21  Attr24  Attr26  Attr
27  Attr30  Attr31  Attr32  Attr33  Attr34  Attr39  Attr40  Attr41  Attr42  Attr43  Attr45  Att
r46  Attr47  Attr49  Attr53  Attr54  Attr57  Attr58  Attr59  Attr60  Attr61  Attr62  Attr63  At
tr64
##   1   3  Attr5  Attr12  Attr13  Attr15  Attr17  Attr19  Attr20  Attr21  Attr24  Attr26  Attr
27  Attr30  Attr31  Attr32  Attr33  Attr34  Attr39  Attr40  Attr41  Attr42  Attr43  Attr45  Att
r46  Attr47  Attr49  Attr53  Attr54  Attr57  Attr58  Attr59  Attr60  Attr61  Attr62  Attr63  At
tr64
##   1   4  Attr5  Attr12  Attr13  Attr15  Attr17  Attr19  Attr20  Attr21  Attr24  Attr26  Attr
27  Attr30  Attr31  Attr32  Attr33  Attr34  Attr39  Attr40  Attr41  Attr42  Attr43  Attr45  Att
r46  Attr47  Attr49  Attr53  Attr54  Attr57  Attr58  Attr59  Attr60  Attr61  Attr62  Attr63  At
tr64
##   1   5  Attr5  Attr12  Attr13  Attr15  Attr17  Attr19  Attr20  Attr21  Attr24  Attr26  Attr
27  Attr30  Attr31  Attr32  Attr33  Attr34  Attr39  Attr40  Attr41  Attr42  Attr43  Attr45  Att
r46  Attr47  Attr49  Attr53  Attr54  Attr57  Attr58  Attr59  Attr60  Attr61  Attr62  Attr63  At
tr64
##   2   1  Attr5  Attr12  Attr13  Attr15  Attr17  Attr19  Attr20  Attr21  Attr24  Attr26  Attr
27  Attr30  Attr31  Attr32  Attr33  Attr34  Attr39  Attr40  Attr41  Attr42  Attr43  Attr45  Att
r46  Attr47  Attr49  Attr53  Attr54  Attr57  Attr58  Attr59  Attr60  Attr61  Attr62  Attr63  At
tr64
##   2   2  Attr5  Attr12  Attr13  Attr15  Attr17  Attr19  Attr20  Attr21  Attr24  Attr26  Attr
27  Attr30  Attr31  Attr32  Attr33  Attr34  Attr39  Attr40  Attr41  Attr42  Attr43  Attr45  Att
r46  Attr47  Attr49  Attr53  Attr54  Attr57  Attr58  Attr59  Attr60  Attr61  Attr62  Attr63  At
tr64
##   2   3  Attr5  Attr12  Attr13  Attr15  Attr17  Attr19  Attr20  Attr21  Attr24  Attr26  Attr
27  Attr30  Attr31  Attr32  Attr33  Attr34  Attr39  Attr40  Attr41  Attr42  Attr43  Attr45  Att
r46  Attr47  Attr49  Attr53  Attr54  Attr57  Attr58  Attr59  Attr60  Attr61  Attr62  Attr63  At
tr64
##   2   4  Attr5  Attr12  Attr13  Attr15  Attr17  Attr19  Attr20  Attr21  Attr24  Attr26  Attr
27  Attr30  Attr31  Attr32  Attr33  Attr34  Attr39  Attr40  Attr41  Attr42  Attr43  Attr45  Att
r46  Attr47  Attr49  Attr53  Attr54  Attr57  Attr58  Attr59  Attr60  Attr61  Attr62  Attr63  At
tr64
##   2   5  Attr5  Attr12  Attr13  Attr15  Attr17  Attr19  Attr20  Attr21  Attr24  Attr26  Attr
27  Attr30  Attr31  Attr32  Attr33  Attr34  Attr39  Attr40  Attr41  Attr42  Attr43  Attr45  Att
r46  Attr47  Attr49  Attr53  Attr54  Attr57  Attr58  Attr59  Attr60  Attr61  Attr62  Attr63  At
tr64
##   3   1  Attr5  Attr12  Attr13  Attr15  Attr17  Attr19  Attr20  Attr21  Attr24  Attr26  Attr
27  Attr30  Attr31  Attr32  Attr33  Attr34  Attr39  Attr40  Attr41  Attr42  Attr43  Attr45  Att
r46  Attr47  Attr49  Attr53  Attr54  Attr57  Attr58  Attr59  Attr60  Attr61  Attr62  Attr63  At
tr64
##   3   2  Attr5  Attr12  Attr13  Attr15  Attr17  Attr19  Attr20  Attr21  Attr24  Attr26  Attr
27  Attr30  Attr31  Attr32  Attr33  Attr34  Attr39  Attr40  Attr41  Attr42  Attr43  Attr45  Att
r46  Attr47  Attr49  Attr53  Attr54  Attr57  Attr58  Attr59  Attr60  Attr61  Attr62  Attr63  At
tr64
##   3   3  Attr5  Attr12  Attr13  Attr15  Attr17  Attr19  Attr20  Attr21  Attr24  Attr26  Attr
27  Attr30  Attr31  Attr32  Attr33  Attr34  Attr39  Attr40  Attr41  Attr42  Attr43  Attr45  Att
r46  Attr47  Attr49  Attr53  Attr54  Attr57  Attr58  Attr59  Attr60  Attr61  Attr62  Attr63  At
tr64
##   3   4  Attr5  Attr12  Attr13  Attr15  Attr17  Attr19  Attr20  Attr21  Attr24  Attr26  Attr
27  Attr30  Attr31  Attr32  Attr33  Attr34  Attr39  Attr40  Attr41  Attr42  Attr43  Attr45  Att
r46  Attr47  Attr49  Attr53  Attr54  Attr57  Attr58  Attr59  Attr60  Attr61  Attr62  Attr63  At
```

```
tr64
##   3   5  Attr5  Attr12  Attr13  Attr15  Attr17  Attr19  Attr20  Attr21  Attr24  Attr26  Attr
27  Attr30  Attr31  Attr32  Attr33  Attr34  Attr39  Attr40  Attr41  Attr42  Attr43  Attr45  Att
r46  Attr47  Attr49  Attr53  Attr54  Attr57  Attr58  Attr59  Attr60  Attr61  Attr62  Attr63  At
tr64
##   4   1  Attr5  Attr12  Attr13  Attr15  Attr17  Attr19  Attr20  Attr21  Attr24  Attr26  Attr
27  Attr30  Attr31  Attr32  Attr33  Attr34  Attr39  Attr40  Attr41  Attr42  Attr43  Attr45  Att
r46  Attr47  Attr49  Attr53  Attr54  Attr57  Attr58  Attr59  Attr60  Attr61  Attr62  Attr63  At
tr64
##   4   2  Attr5  Attr12  Attr13  Attr15  Attr17  Attr19  Attr20  Attr21  Attr24  Attr26  Attr
27  Attr30  Attr31  Attr32  Attr33  Attr34  Attr39  Attr40  Attr41  Attr42  Attr43  Attr45  Att
r46  Attr47  Attr49  Attr53  Attr54  Attr57  Attr58  Attr59  Attr60  Attr61  Attr62  Attr63  At
tr64
##   4   3  Attr5  Attr12  Attr13  Attr15  Attr17  Attr19  Attr20  Attr21  Attr24  Attr26  Attr
27  Attr30  Attr31  Attr32  Attr33  Attr34  Attr39  Attr40  Attr41  Attr42  Attr43  Attr45  Att
r46  Attr47  Attr49  Attr53  Attr54  Attr57  Attr58  Attr59  Attr60  Attr61  Attr62  Attr63  At
tr64
##   4   4  Attr5  Attr12  Attr13  Attr15  Attr17  Attr19  Attr20  Attr21  Attr24  Attr26  Attr
27  Attr30  Attr31  Attr32  Attr33  Attr34  Attr39  Attr40  Attr41  Attr42  Attr43  Attr45  Att
r46  Attr47  Attr49  Attr53  Attr54  Attr57  Attr58  Attr59  Attr60  Attr61  Attr62  Attr63  At
tr64
##   4   5  Attr5  Attr12  Attr13  Attr15  Attr17  Attr19  Attr20  Attr21  Attr24  Attr26  Attr
27  Attr30  Attr31  Attr32  Attr33  Attr34  Attr39  Attr40  Attr41  Attr42  Attr43  Attr45  Att
r46  Attr47  Attr49  Attr53  Attr54  Attr57  Attr58  Attr59  Attr60  Attr61  Attr62  Attr63  At
tr64
##   5   1  Attr5  Attr12  Attr13  Attr15  Attr17  Attr19  Attr20  Attr21  Attr24  Attr26  Attr
27  Attr30  Attr31  Attr32  Attr33  Attr34  Attr39  Attr40  Attr41  Attr42  Attr43  Attr45  Att
r46  Attr47  Attr49  Attr53  Attr54  Attr57  Attr58  Attr59  Attr60  Attr61  Attr62  Attr63  At
tr64
##   5   2  Attr5  Attr12  Attr13  Attr15  Attr17  Attr19  Attr20  Attr21  Attr24  Attr26  Attr
27  Attr30  Attr31  Attr32  Attr33  Attr34  Attr39  Attr40  Attr41  Attr42  Attr43  Attr45  Att
r46  Attr47  Attr49  Attr53  Attr54  Attr57  Attr58  Attr59  Attr60  Attr61  Attr62  Attr63  At
tr64
##   5   3  Attr5  Attr12  Attr13  Attr15  Attr17  Attr19  Attr20  Attr21  Attr24  Attr26  Attr
27  Attr30  Attr31  Attr32  Attr33  Attr34  Attr39  Attr40  Attr41  Attr42  Attr43  Attr45  Att
r46  Attr47  Attr49  Attr53  Attr54  Attr57  Attr58  Attr59  Attr60  Attr61  Attr62  Attr63  At
tr64
##   5   4  Attr5  Attr12  Attr13  Attr15  Attr17  Attr19  Attr20  Attr21  Attr24  Attr26  Attr
27  Attr30  Attr31  Attr32  Attr33  Attr34  Attr39  Attr40  Attr41  Attr42  Attr43  Attr45  Att
r46  Attr47  Attr49  Attr53  Attr54  Attr57  Attr58  Attr59  Attr60  Attr61  Attr62  Attr63  At
tr64
##   5   5  Attr5  Attr12  Attr13  Attr15  Attr17  Attr19  Attr20  Attr21  Attr24  Attr26  Attr
27  Attr30  Attr31  Attr32  Attr33  Attr34  Attr39  Attr40  Attr41  Attr42  Attr43  Attr45  Att
r46  Attr47  Attr49  Attr53  Attr54  Attr57  Attr58  Attr59  Attr60  Attr61  Attr62  Attr63  At
tr64
```

```
## Warning: Number of logged events: 875
```

```
complete.data = complete(imp)

for(i in 1:ncol(complete.data)){
  if(sum(is.na(complete.data[,i] == T))){
    cat("Attr", i, "has", sum(is.na(complete.data[,i])), "NAs", "₩n")
  }
}
```

모두 다 채워졌다!

```
imp$method
```

```
##  Attr1  Attr2  Attr3  Attr5  Attr6  Attr7  Attr9 Attr10 Attr12 Attr13 Attr15
##     ""     ""     ""  "pmm"     ""     ""     ""     ""  "pmm"  "pmm"  "pmm"
## Attr17 Attr18 Attr19 Attr20 Attr21 Attr22 Attr24 Attr25 Attr26 Attr27 Attr29
##  "pmm"     ""  "pmm"  "pmm"  "pmm"     ""  "pmm"     ""  "pmm"  "pmm"     ""
## Attr30 Attr31 Attr32 Attr33 Attr34 Attr35 Attr36 Attr38 Attr39 Attr40 Attr41
##  "pmm"  "pmm"  "pmm"  "pmm"  "pmm"     ""     ""     ""  "pmm"  "pmm"  "pmm"
## Attr42 Attr43 Attr45 Attr46 Attr47 Attr48 Attr49 Attr51 Attr53 Attr54 Attr55
##  "pmm"  "pmm"  "pmm"  "pmm"  "pmm"     ""  "pmm"     ""  "pmm"  "pmm"     ""
## Attr57 Attr58 Attr59 Attr60 Attr61 Attr62 Attr63 Attr64  class
##  "pmm"  "pmm"  "pmm"  "pmm"  "pmm"  "pmm"  "pmm"  "pmm"     ""
```

PMM(Predictive mean matching) 방식으로 Imputation을 진행했다.

저장하자

```
write.csv(complete.data, "imputed_data_mice.csv", row.names = FALSE)
```

3-2) impute 함수의 KNN imputation 활용

```
knn_imputed = impute.knn(as.matrix(data))
```

```
## Cluster size 6855 broken into 6815 40
## Cluster size 6815 broken into 6306 509
## Cluster size 6306 broken into 6301 5
## Cluster size 6301 broken into 1316 4985
## Done cluster 1316
## Cluster size 4985 broken into 54 4931
## Done cluster 54
## Cluster size 4931 broken into 4667 264
## Cluster size 4667 broken into 4507 160
## Cluster size 4507 broken into 3225 1282
## Cluster size 3225 broken into 621 2604
## Done cluster 621
## Cluster size 2604 broken into 1820 784
## Cluster size 1820 broken into 1807 13
## Cluster size 1807 broken into 5 1802
## Done cluster 5
## Cluster size 1802 broken into 841 961
## Done cluster 841
## Done cluster 961
## Done cluster 1802
## Done cluster 1807
## Done cluster 13
## Done cluster 1820
## Done cluster 784
## Done cluster 2604
## Done cluster 3225
## Done cluster 1282
## Done cluster 4507
## Done cluster 160
## Done cluster 4667
## Done cluster 264
## Done cluster 4931
## Done cluster 4985
## Done cluster 6301
## Done cluster 5
## Done cluster 6306
## Done cluster 509
## Done cluster 6815
## Done cluster 40
```

저장하자

```
complete_data_knn = knn_imputed$data

write.csv(complete_data_knn, "imputed_data_knn.csv", row.names = FALSE)
```

3-3) Mean Imputation

가장 간단하게 그 칼럼의 mean 값으로 NA를 채워넣는 방법이다.

```
complete_data_mean = as.matrix(data)

for(i in 1:ncol(complete_data_mean)) {
  complete_data_mean[ , i][is.na(complete_data_mean[ , i])] <- mean(complete_data_mean[ , i], n
a.rm = TRUE)
}

complete_data_mean = data.frame(complete_data_mean)

write.csv(complete_data_mean, "imputed_data_mean.csv", row.names = FALSE)
```