Final_programming

Biswajit Chowdhury

03/12/2022

Description:

This is a primer for missing value imputation. Here, I used the 'heart data' from UCI depository. The dataset does not included missing values. Therefore, I introduuced 20% missing value using introduced missing values of the total observation using the prodNA function from the missForest package. The prodNA function incorporates missing values completely at random. Therefore, the characteristics of missing values in the given dataset are missing completely at random (MCAR). The missing values were imputed by four diffierent ways using Multiple imputation using chained equations (MICE).

Dataset 1: Imputed all variables using default MICE function. In this function, numerical variables were imputed by predictive mean matching (PMM) and categorical variables either by logreg, or polyreg. This dataset has been defined as 'MICE' in this analysis.

Dataset 2: Imputed all variables by PMM. This is a semi-parametric imputation approach, which draws imputed values from an observed empirical distribution. This has been defined as "PMM" in this article.

Dataset 3: Changed numerical variable with mean and kept categorical as dataset one and defined this as 'Mean'.

Dataset 4: Introduced missing values only in the numerical variables (age, resting blood pressure, cholesterol level, max_heart rate and oldpeak) and keep others unchanged (as train data). Then replaced the missing values only for five numerical variables with mean imputation and labeled it as 'mean_numeric'.

Load library

Reading data file

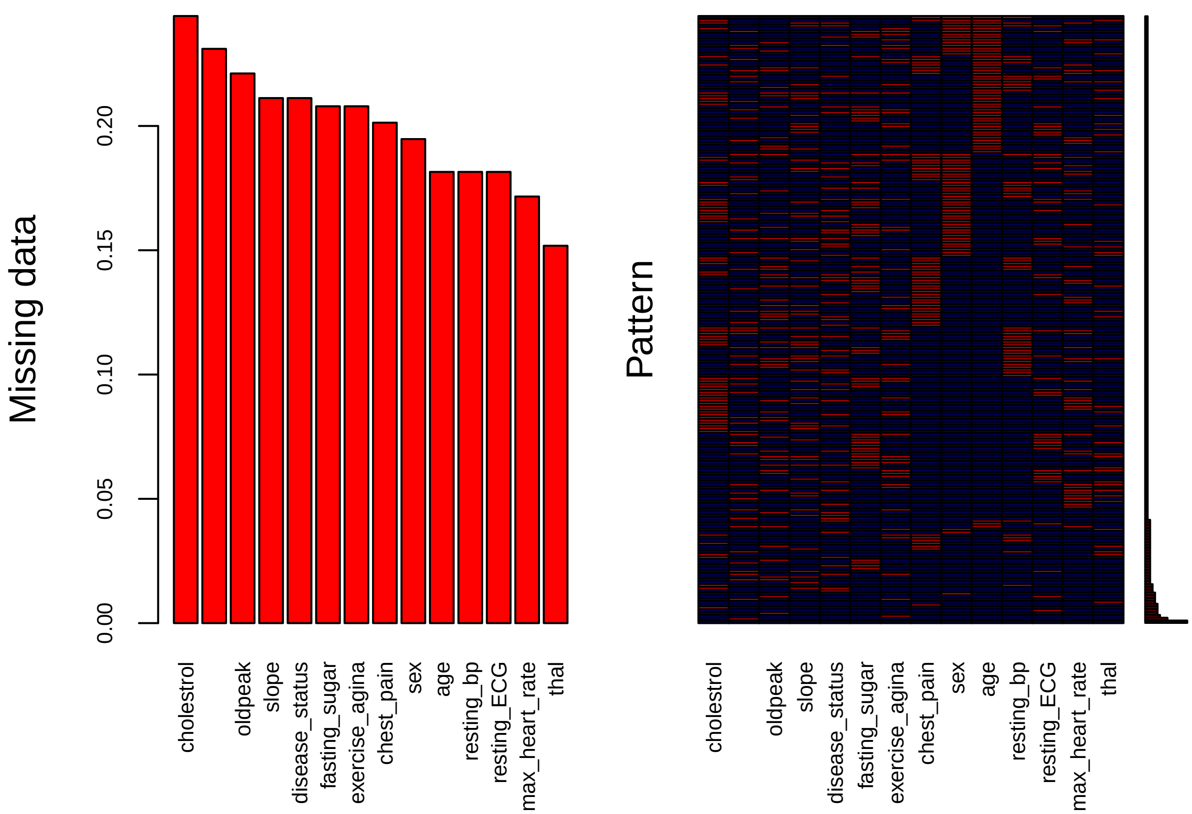
```
##   age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal
## 1  63   1  3      145  233   1       0     150     0     2.3     0  0    1
## 2  37   1  2      130  250   0       1     187     0     3.5     0  0    2
## 3  41   0  1      130  204   0       0     172     0     1.4     2  0    2
##   target
## 1      1
## 2      1
## 3      1
```

```
## 'data.frame':    303 obs. of  14 variables:
##  $ age               : int  63 37 41 56 57 57 56 44 52 57 ...
##  $ sex               : int  1 1 0 1 0 1 0 1 1 1 ...
##  $ chest_pain        : int  3 2 1 1 0 0 1 1 2 2 ...
```

```
##  $ resting_bp           : int  145 130 130 120 120 140 140 120 172 150 ...
##  $ cholestrol           : int  233 250 204 236 354 192 294 263 199 168 ...
##  $ fasting_sugar        : int  1 0 0 0 0 0 0 0 1 0 ...
##  $ resting_ECG          : int  0 1 0 1 1 1 0 1 1 1 ...
##  $ max_heart_rate       : int  150 187 172 178 163 148 153 173 162 174 ...
##  $ exercise_agina       : int  0 0 0 0 1 0 0 0 0 0 ...
##  $ oldpeak              : num  2.3 3.5 1.4 0.8 0.6 0.4 1.3 0 0.5 1.6 ...
##  $ slope                : int  0 0 2 2 2 1 1 2 2 2 ...
##  $ number_major_vessels: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ thal                 : int  1 2 2 2 2 1 2 3 3 2 ...
##  $ disease_status       : int  1 1 1 1 1 1 1 1 1 1 ...
```

Preprocess the variables

Create missing values in the data set




```
##
##  Variables sorted by number of missings:
##              Variable     Count
##            cholestrol 0.2442244
##  number_major_vessels 0.2310231
##               oldpeak 0.2211221
##                 slope 0.2112211
```

```
##        disease_status 0.2112211
##         fasting_sugar 0.2079208
##        exercise_agina 0.2079208
##            chest_pain 0.2013201
##                   sex 0.1947195
##                   age 0.1815182
##            resting_bp 0.1815182
##           resting_ECG 0.1815182
##        max_heart_rate 0.1716172
##                  thal 0.1518152
```

Figure 1. Aggregation plot for missing data on the train data set. The missing values were introduced using the prodNA function from the miss Forest package. The visualization of the missing values patterns was generated using the VIM package in R. (A) represents the proportion of the missing and (B) distribution patterns in each variable. The red in graph representing missing values along with observed values (blue).

1. Multiple imputation

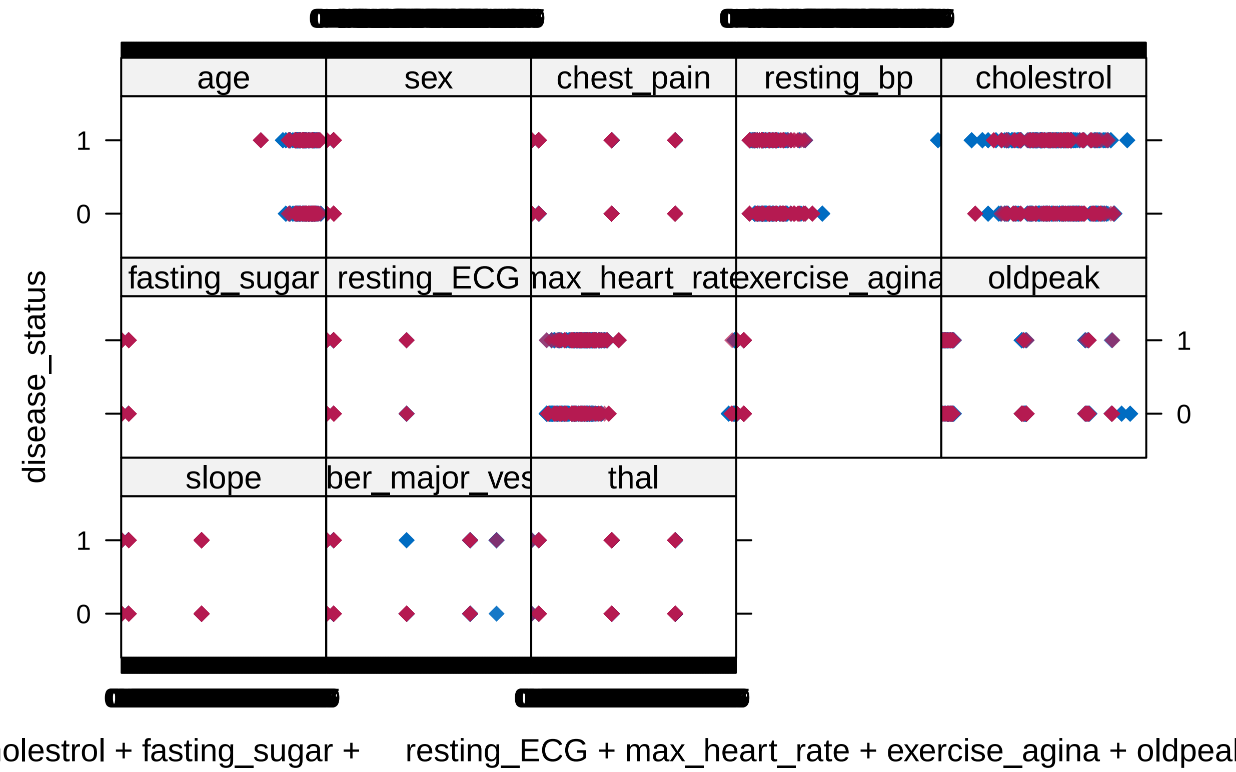
impute missing values using multivariate imputation by chained equations (MICE)

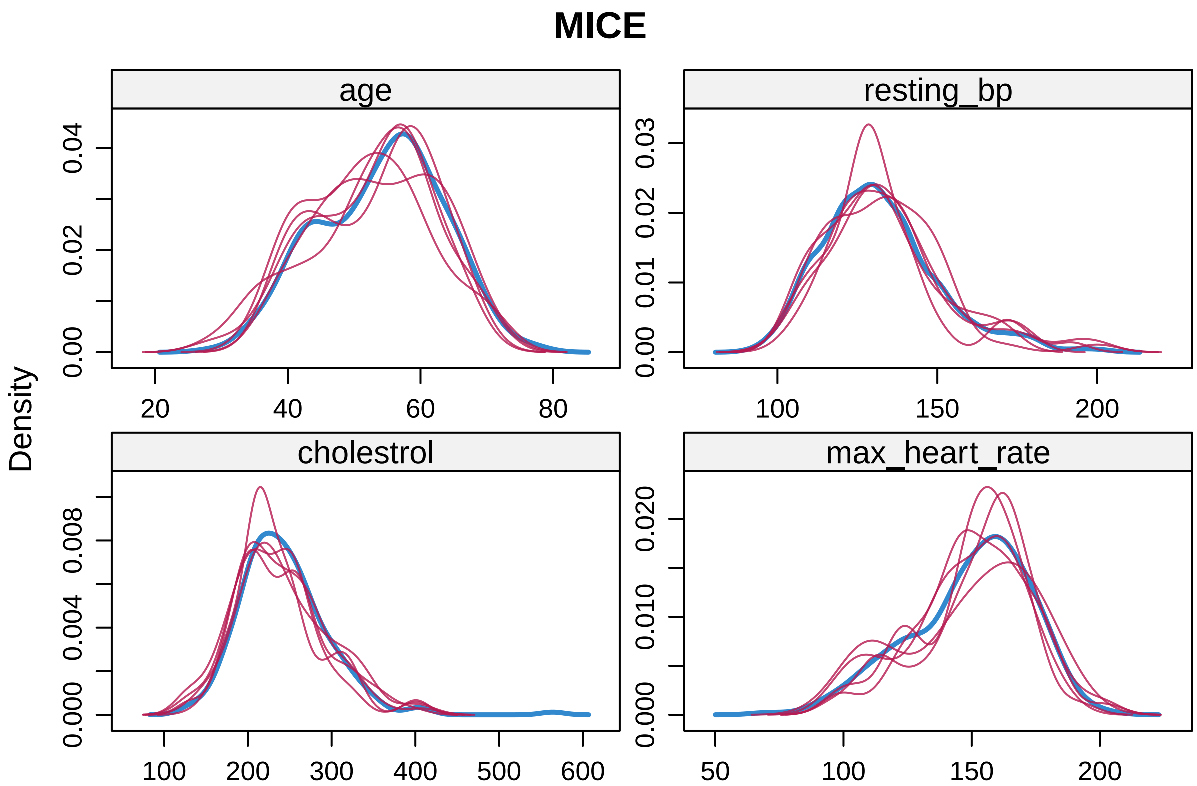
```
## Class: mids
## Number of multiple imputations:  5
## Imputation methods:
##                  age                  sex           chest_pain
##                "pmm"             "logreg"            "polyreg"
##           resting_bp           cholestrol        fasting_sugar
##                "pmm"                "pmm"             "logreg"
##          resting_ECG       max_heart_rate       exercise_agina
##            "polyreg"                "pmm"             "logreg"
##              oldpeak                slope number_major_vessels
##                "pmm"            "polyreg"            "polyreg"
##                 thal       disease_status
##            "polyreg"             "logreg"
## PredictorMatrix:
##               age sex chest_pain resting_bp cholestrol fasting_sugar
## age             0   1          1          1          1             1
## sex             1   0          1          1          1             1
## chest_pain      1   1          0          1          1             1
## resting_bp      1   1          1          0          1             1
## cholestrol      1   1          1          1          0             1
## fasting_sugar   1   1          1          1          1             0
##               resting_ECG max_heart_rate exercise_agina oldpeak slope
## age                     1              1              1       1     1
## sex                     1              1              1       1     1
## chest_pain              1              1              1       1     1
## resting_bp              1              1              1       1     1
## cholestrol              1              1              1       1     1
## fasting_sugar           1              1              1       1     1
##               number_major_vessels thal disease_status
## age                              1    1              1
## sex                              1    1              1
```

```
## chest_pain                    1    1              1
## resting_bp                    1    1              1
## cholestrol                    1    1              1
## fasting_sugar                 1    1              1
```

```
## [1] 1515   16
```

#Inspecting the distribution of original and imputed data

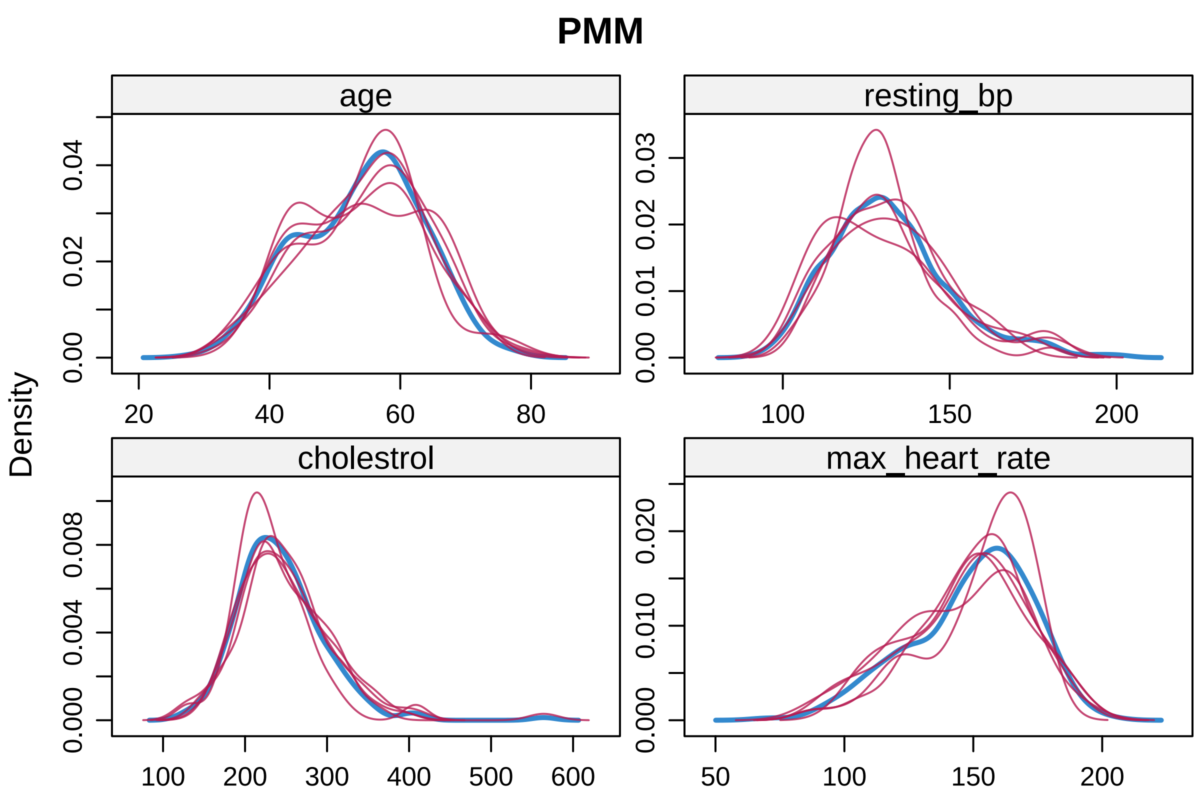




2. Multiple imputation with PMM

```
## Class: mids
## Number of multiple imputations:  5
## Imputation methods:
##                  age                  sex           chest_pain
##                "pmm"                "pmm"                "pmm"
##           resting_bp           cholestrol        fasting_sugar
##                "pmm"                "pmm"                "pmm"
##          resting_ECG       max_heart_rate       exercise_agina
##                "pmm"                "pmm"                "pmm"
##              oldpeak                slope number_major_vessels
##                "pmm"                "pmm"                "pmm"
##                 thal       disease_status
##                "pmm"                "pmm"
## PredictorMatrix:
##                age sex chest_pain resting_bp cholestrol fasting_sugar
## age              0   1          1          1          1             1
## sex              1   0          1          1          1             1
## chest_pain       1   1          0          1          1             1
## resting_bp       1   1          1          0          1             1
## cholestrol       1   1          1          1          0             1
## fasting_sugar    1   1          1          1          1             0
##                resting_ECG max_heart_rate exercise_agina oldpeak slope
```

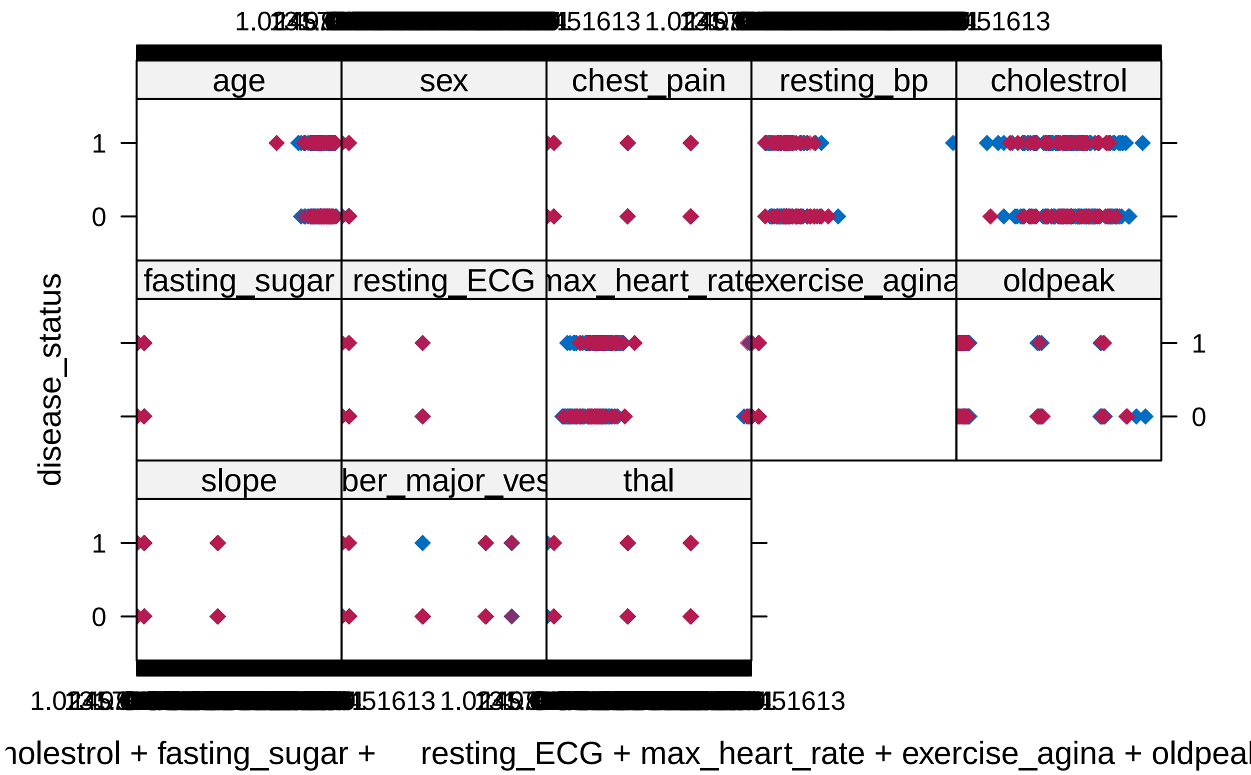
```
## age                   1              1              1       1     1
## sex                   1              1              1       1     1
## chest_pain            1              1              1       1     1
## resting_bp            1              1              1       1     1
## cholestrol            1              1              1       1     1
## fasting_sugar         1              1              1       1     1
##                number_major_vessels thal disease_status
## age                               1    1              1
## sex                               1    1              1
## chest_pain                        1    1              1
## resting_bp                        1    1              1
## cholestrol                        1    1              1
## fasting_sugar                     1    1              1
```

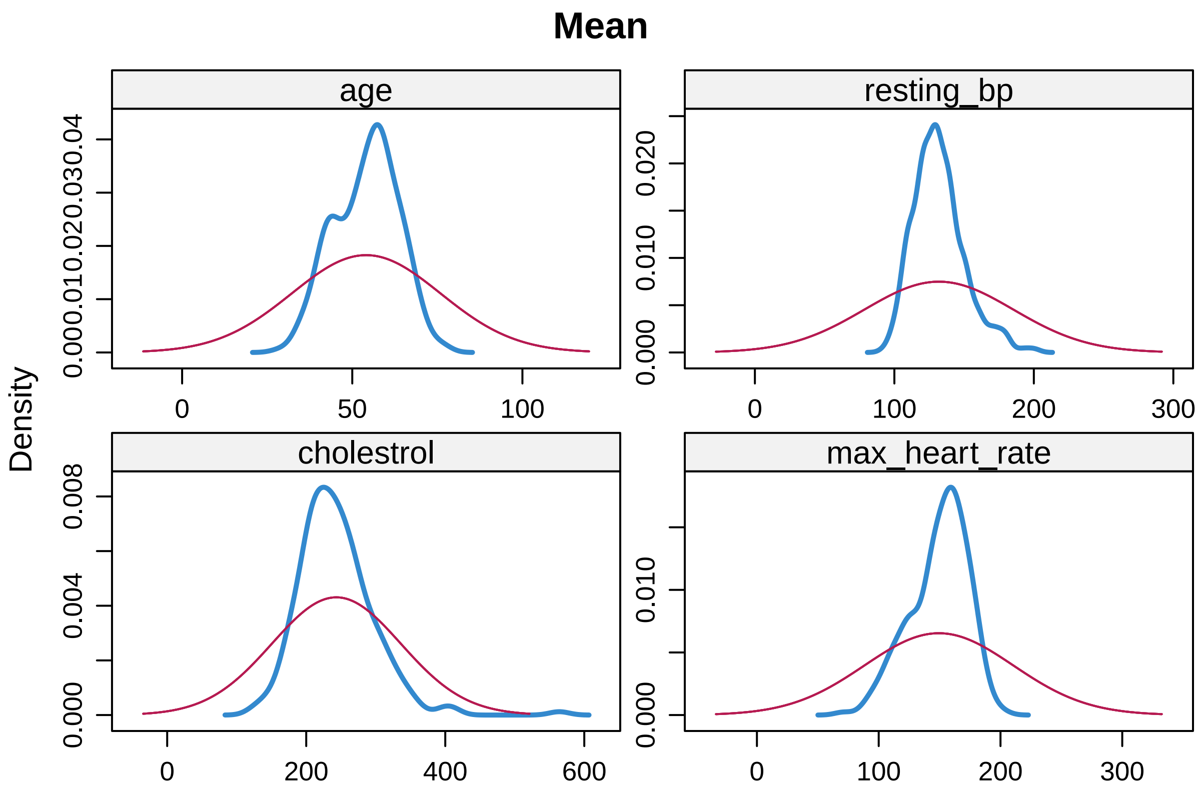


3. Multiple imputation with Mean

```
## Class: mids
## Number of multiple imputations:  5
## Imputation methods:
##                 age                 sex          chest_pain
##              "mean"            "logreg"           "polyreg"
##          resting_bp          cholestrol       fasting_sugar
##              "mean"              "mean"            "logreg"
##         resting_ECG      max_heart_rate      exercise_agina
```

```
##            "polyreg"               "mean"             "logreg"
##              oldpeak                slope number_major_vessels
##               "mean"            "polyreg"            "polyreg"
##                 thal       disease_status
##            "polyreg"             "logreg"
## PredictorMatrix:
##               age sex chest_pain resting_bp cholestrol fasting_sugar
## age             0   1          1          1          1             1
## sex             1   0          1          1          1             1
## chest_pain      1   1          0          1          1             1
## resting_bp      1   1          1          0          1             1
## cholestrol      1   1          1          1          0             1
## fasting_sugar   1   1          1          1          1             0
##               resting_ECG max_heart_rate exercise_agina oldpeak slope
## age                     1              1              1       1     1
## sex                     1              1              1       1     1
## chest_pain              1              1              1       1     1
## resting_bp              1              1              1       1     1
## cholestrol              1              1              1       1     1
## fasting_sugar           1              1              1       1     1
##               number_major_vessels thal disease_status
## age                              1    1              1
## sex                              1    1              1
## chest_pain                       1    1              1
## resting_bp                       1    1              1
## cholestrol                       1    1              1
## fasting_sugar                    1    1              1
```


age
sex
chest_pain
resting_bp
cholestrol
fasting_sugar
resting_ECG
max_heart_rate
exercise_agina
oldpeak
slope
ber_major_ves
thal
disease_status
1
0
olestrol + fasting_sugar + resting_ECG + max_heart_rate + exercise_agina + oldpeak



4. Adding missing values only in numerical variables

```
##       age          resting_bp      cholestrol     max_heart_rate
##  Min.   :29.00   Min.   : 94.0   Min.   :131.0   Min.   : 71.0
##  1st Qu.:47.00   1st Qu.:120.0   1st Qu.:212.0   1st Qu.:135.5
##  Median :55.00   Median :130.0   Median :243.0   Median :154.0
##  Mean   :54.35   Mean   :131.5   Mean   :248.2   Mean   :149.7
##  3rd Qu.:61.00   3rd Qu.:140.0   3rd Qu.:277.0   3rd Qu.:166.0
##  Max.   :77.00   Max.   :200.0   Max.   :564.0   Max.   :202.0
##  NA's   :66      NA's   :68      NA's   :62      NA's   :59
##     oldpeak
##  Min.   :0.000
##  1st Qu.:0.000
##  Median :0.800
##  Mean   :1.034
##  3rd Qu.:1.600
##  Max.   :6.200
##  NA's   :48
```

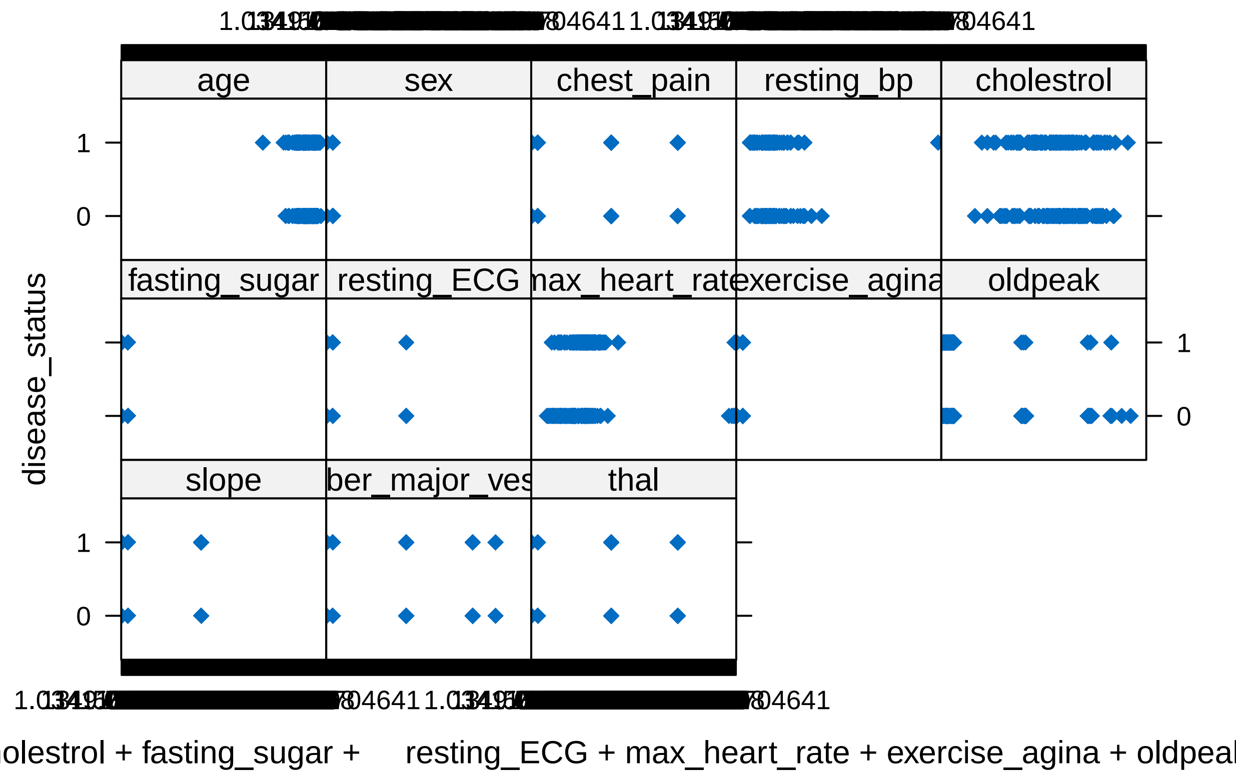
Impute numerical missing values using mean imputation

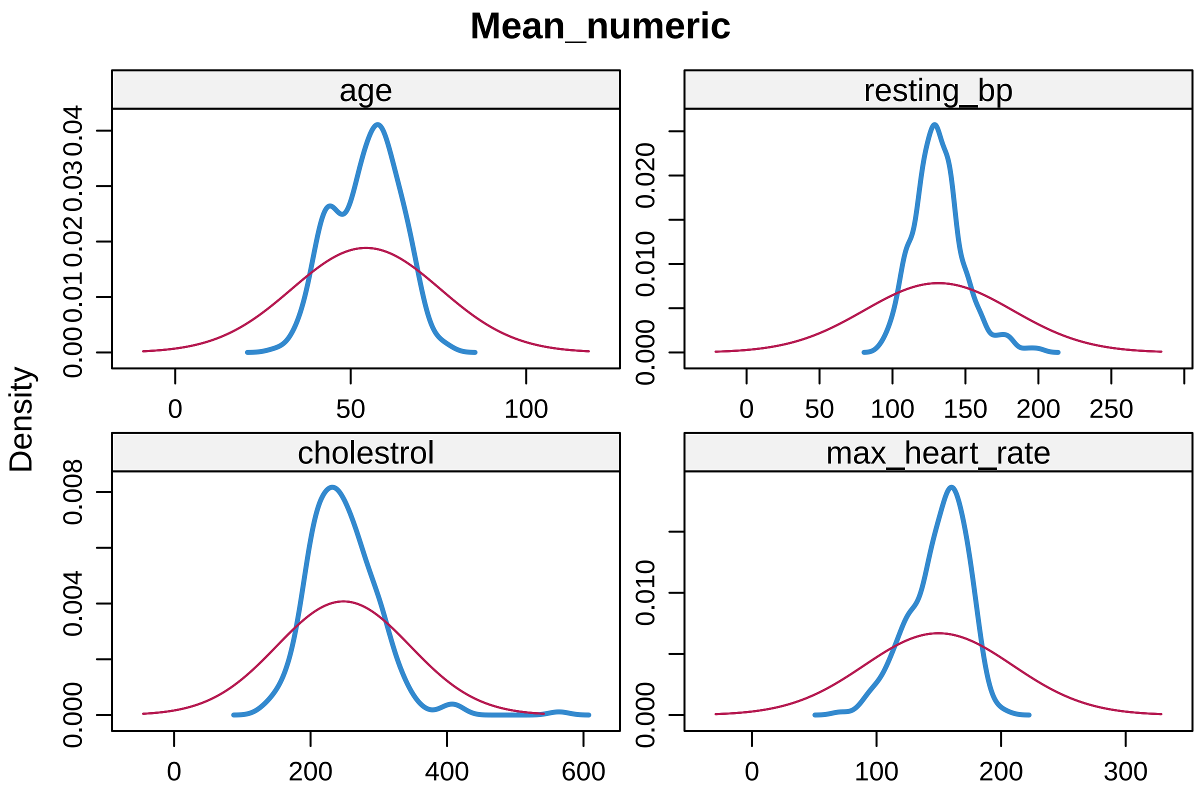
```
## Class: mids
## Number of multiple imputations:  5
```

```
## Imputation methods:
##                  age                  sex           chest_pain
##               "mean"                   ""                   ""
##           resting_bp           cholestrol        fasting_sugar
##               "mean"               "mean"                   ""
##          resting_ECG        max_heart_rate       exercise_agina
##                   ""               "mean"                   ""
##              oldpeak                slope number_major_vessels
##               "mean"                   ""                   ""
##                 thal       disease_status
##                   ""                   ""
## PredictorMatrix:
##               age sex chest_pain resting_bp cholestrol fasting_sugar
## age             0   1          1          1          1             1
## sex             1   0          1          1          1             1
## chest_pain      1   1          0          1          1             1
## resting_bp      1   1          1          0          1             1
## cholestrol      1   1          1          1          0             1
## fasting_sugar   1   1          1          1          1             0
##               resting_ECG max_heart_rate exercise_agina oldpeak slope
## age                     1              1              1       1     1
## sex                     1              1              1       1     1
## chest_pain              1              1              1       1     1
## resting_bp              1              1              1       1     1
## cholestrol              1              1              1       1     1
## fasting_sugar           1              1              1       1     1
##               number_major_vessels thal disease_status
## age                              1    1              1
## sex                              1    1              1
## chest_pain                       1    1              1
## resting_bp                       1    1              1
## cholestrol                       1    1              1
## fasting_sugar                    1    1              1
```

```
##       .imp        .id           age          sex       chest_pain   resting_bp   
##  Min.   :1   Min.   :  1   Min.   :29.00   0: 480   0:715      Min.   : 94.0  
##  1st Qu.:2   1st Qu.: 76   1st Qu.:50.00   1:1035   1:250      1st Qu.:123.0  
##  Median :3   Median :152   Median :54.35            2:435      Median :131.5  
##  Mean   :3   Mean   :152   Mean   :54.35            3:115      Mean   :131.5  
##  3rd Qu.:4   3rd Qu.:228   3rd Qu.:59.00                       3rd Qu.:138.0  
##  Max.   :5   Max.   :303   Max.   :77.00                       Max.   :200.0  
##    cholestrol    fasting_sugar resting_ECG max_heart_rate  exercise_agina
##  Min.   :131.0   0:1290        0:735       Min.   : 71.0   0:1020        
##  1st Qu.:221.0   1: 225        1:760       1st Qu.:142.0   1: 495        
##  Median :248.2                 2: 20       Median :149.7                 
##  Mean   :248.2                             Mean   :149.7                 
##  3rd Qu.:268.0                             3rd Qu.:163.0                 
##  Max.   :564.0                             Max.   :202.0                 
##     oldpeak      slope   number_major_vessels thal    disease_status
##  Min.   :0.000   0:105   0:875                0: 10   0:690         
##  1st Qu.:0.000   1:700   1:325                1: 90   1:825         
##  Median :1.034   2:710   2:190                2:830                 
##  Mean   :1.034           3:100                3:585                 
##  3rd Qu.:1.400           4: 25                                      
```

```
##   Max.    :6.200
```


age
sex
chest_pain
resting_bp
cholestrol
fasting_sugar
resting_ECG
max_heart_rate
exercise_agina
oldpeak
slope
thal
disease_status
1
0
holestrol + fasting_sugar +
resting_ECG + max_heart_rate + exercise_agina + oldpeak



Mean_numeric
age
resting_bp
cholestrol
max_heart_rate
Density
0.000
0.010
0.020
0.030
0.04
0.004
0.008
0
50
100
150
200
250
400
600
300

Summary table for original and imputed data using gtsummary

Characteristic	**Mean**, N = 303	**Original**, N = 303	**PMM**, N = 303
age	54 (8)	54 (9)	54 (9)
sex			
0	303 (29%)	303 (32%)	303 (32%)
1	303 (71%)	303 (68%)	303 (68%)
chest_pain			
0	303 (48%)	303 (47%)	303 (48%)
1	303 (17%)	303 (17%)	303 (17%)
2	303 (27%)	303 (29%)	303 (28%)
3	303 (7.3%)	303 (7.6%)	303 (7.3%)
resting_bp	132 (16)	132 (18)	132 (18)
cholestrol	243 (46)	246 (52)	245 (53)
fasting_sugar			
0	303 (84%)	303 (85%)	303 (84%)
1	303 (16%)	303 (15%)	303 (16%)
resting_ECG			
0	303 (48%)	303 (49%)	303 (49%)
1	303 (49%)	303 (50%)	303 (50%)
2	303 (2.6%)	303 (1.3%)	303 (1.0%)
max_heart_rate	149 (21)	150 (23)	149 (23)
exercise_agina			
0	303 (64%)	303 (67%)	303 (65%)
1	303 (36%)	303 (33%)	303 (35%)
oldpeak	1.02 (1.04)	1.04 (1.16)	1.00 (1.17)
slope			
0	303 (8.3%)	303 (6.9%)	303 (7.9%)
1	303 (47%)	303 (46%)	303 (48%)
2	303 (45%)	303 (47%)	303 (45%)
number_major_vessels			
0	303 (58%)	303 (58%)	303 (59%)
1	303 (22%)	303 (21%)	303 (23%)
2	303 (12%)	303 (13%)	303 (11%)
3	303 (6.6%)	303 (6.6%)	303 (5.9%)
4	303 (1.7%)	303 (1.7%)	303 (1.0%)
thal			
0	303 (1.3%)	303 (0.7%)	303 (1.0%)
1	303 (7.3%)	303 (5.9%)	303 (5.9%)
2	303 (53%)	303 (55%)	303 (54%)
3	303 (38%)	303 (39%)	303 (39%)
disease_status			
0	303 (47%)	303 (46%)	303 (48%)
1	303 (53%)	303 (54%)	303 (52%)