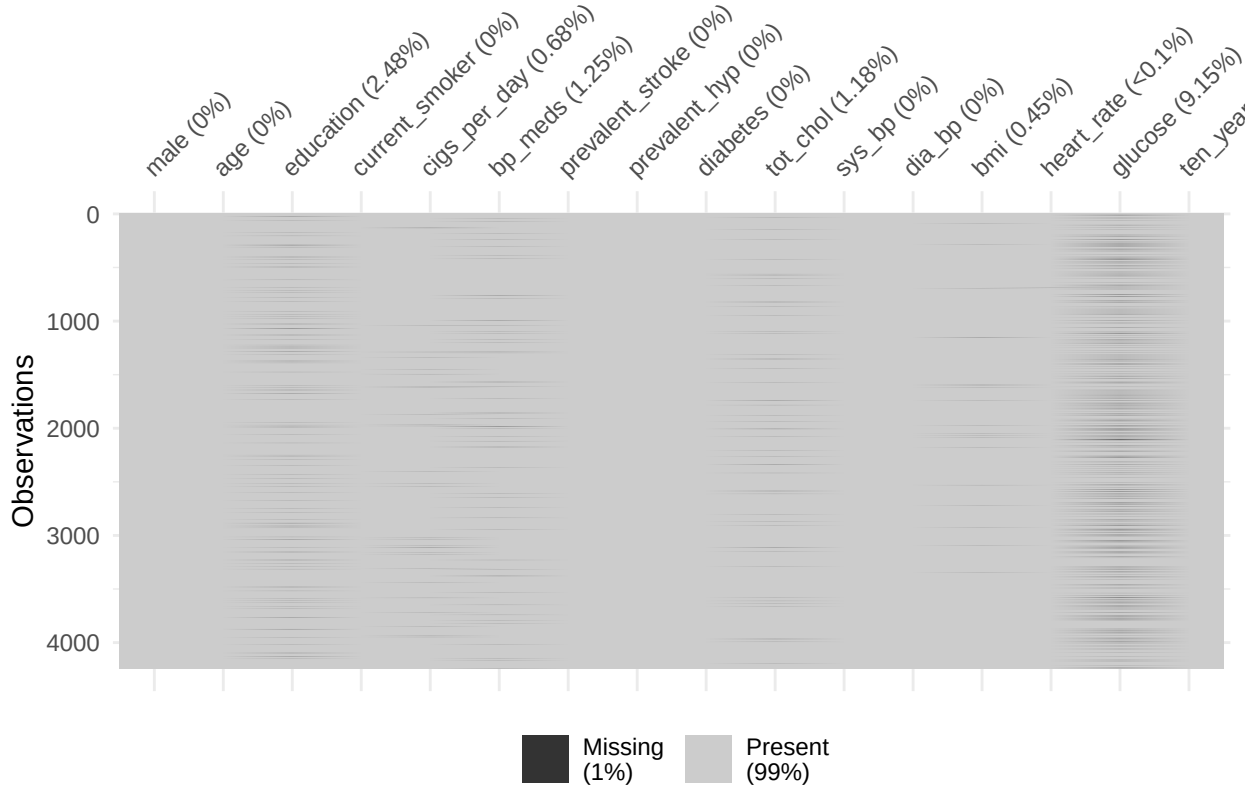
Missing Data Imputation

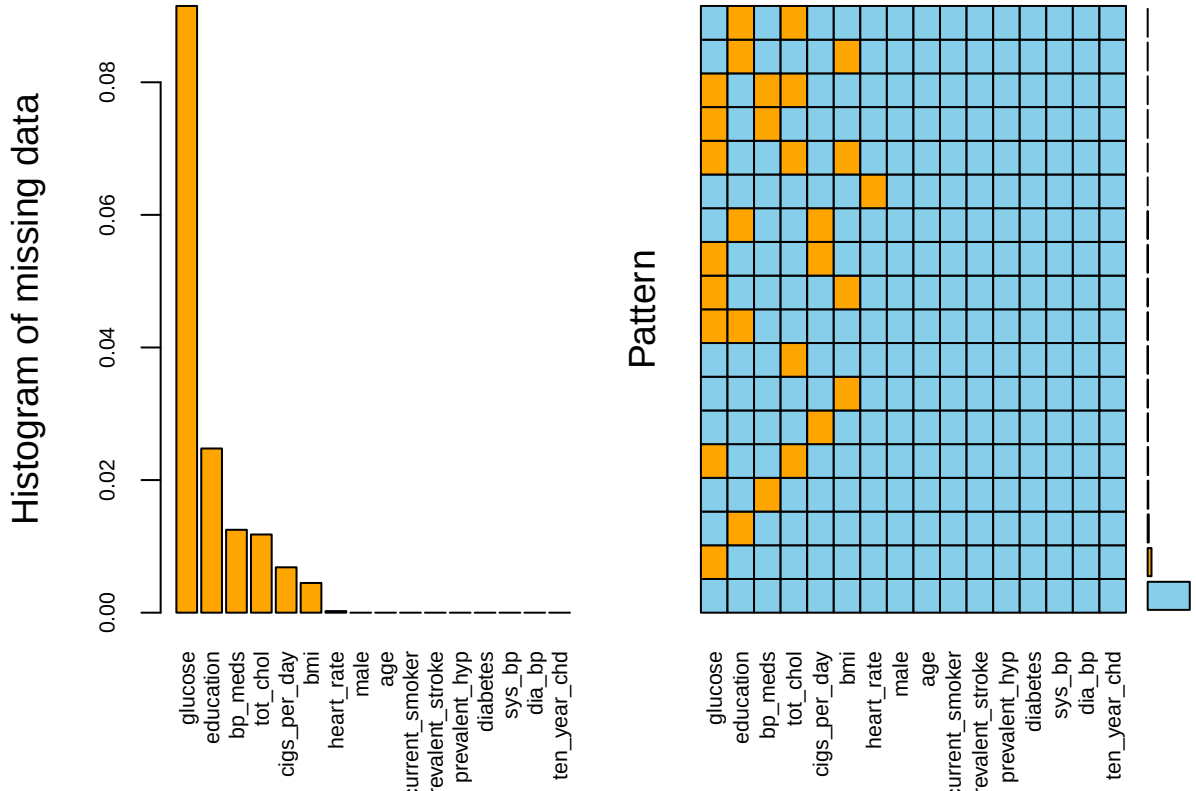
Hun

2022-04-28

Visualizing missing data



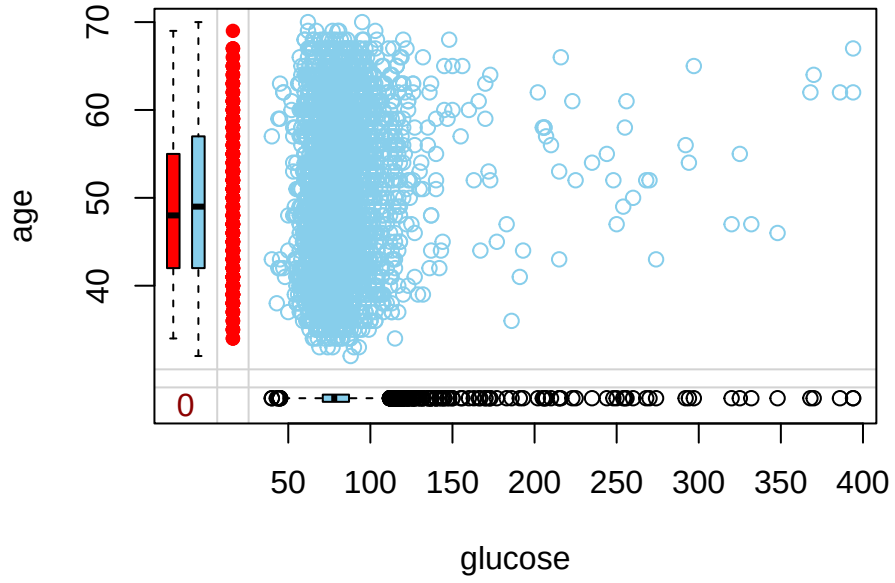
Patterns of missing data

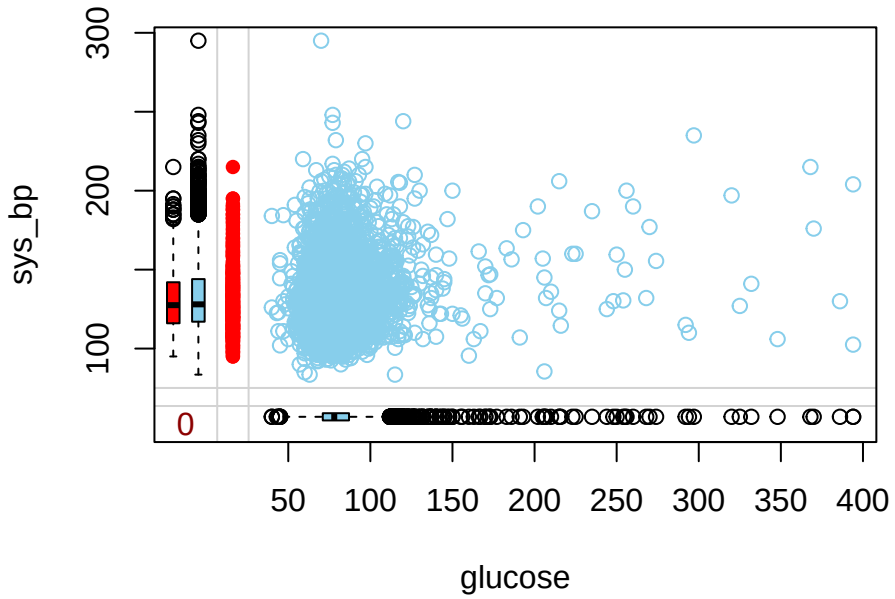



```
##
##  Variables sorted by number of missings:
##           Variable        Count
##            glucose 0.0915094340
##          education 0.0247641509
##            bp_meds 0.0125000000
##           tot_chol 0.0117924528
##      cigs_per_day 0.0068396226
##                bmi 0.0044811321
##         heart_rate 0.0002358491
##               male 0.0000000000
##                age 0.0000000000
##    current_smoker 0.0000000000
##  prevalent_stroke 0.0000000000
##     prevalent_hyp 0.0000000000
##           diabetes 0.0000000000
##             sys_bp 0.0000000000
##             dia_bp 0.0000000000
##      ten_year_chd 0.0000000000
```

Checking the assumption of missing completely at random (MCAR)

If our assumption of MCAR data is correct, then we expect the red and blue box plots to be very similar.



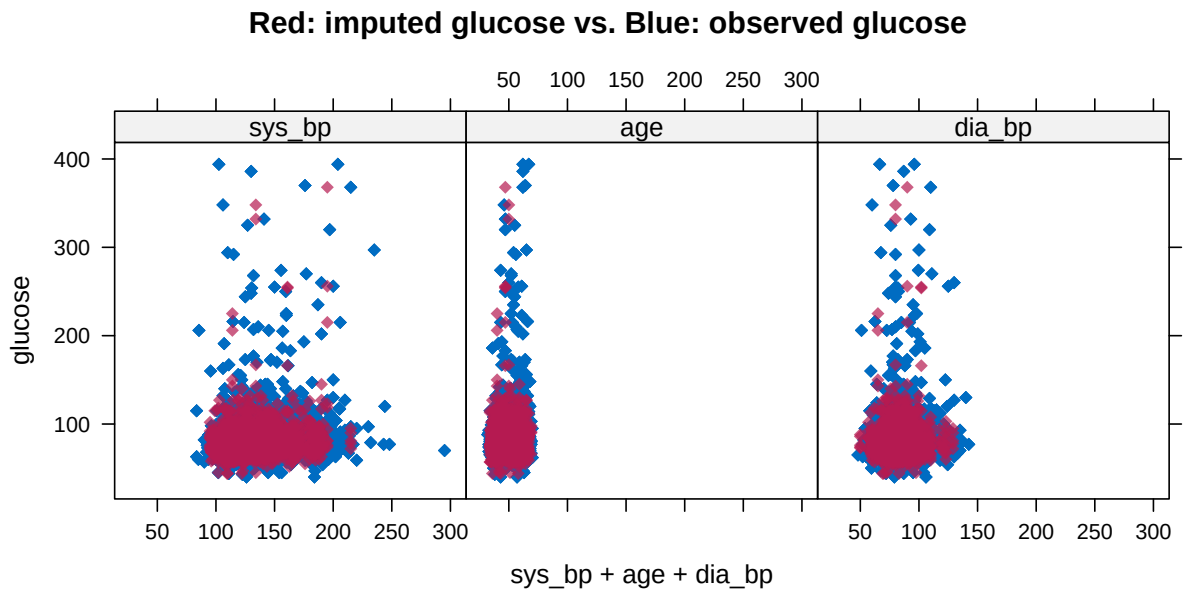


Selecting variables to impute data

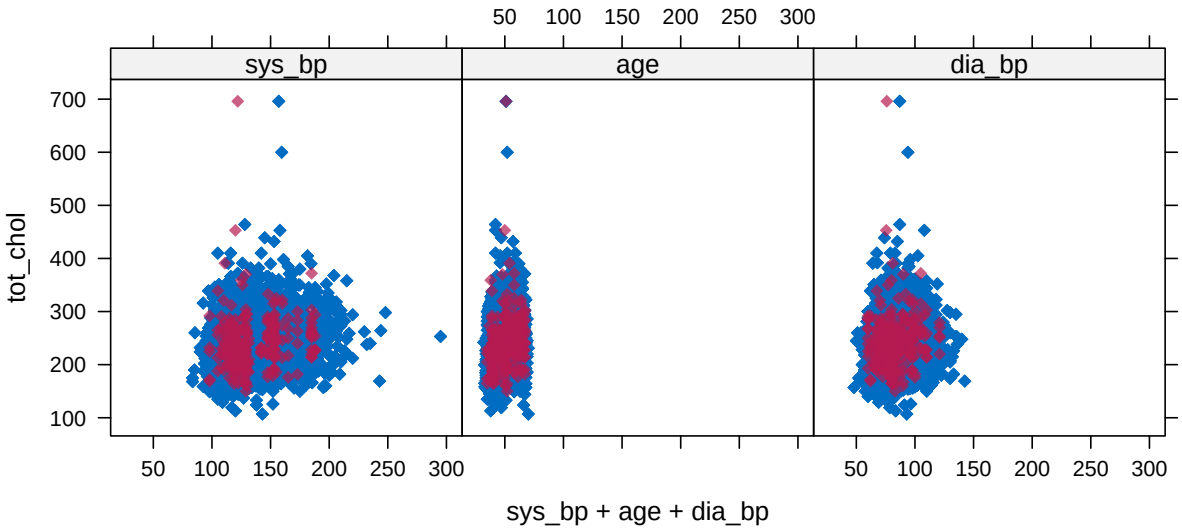
I selected *glucose* and *tot_chol* because they both are continuous and have similar patterns.

Using Predictive Mean Matching (PMM) to impute missing data

```
##
##  iter imp variable
##   1   1  tot_chol  glucose
##   1   2  tot_chol  glucose
##   1   3  tot_chol  glucose
##   1   4  tot_chol  glucose
##   1   5  tot_chol  glucose
##   2   1  tot_chol  glucose
##   2   2  tot_chol  glucose
##   2   3  tot_chol  glucose
##   2   4  tot_chol  glucose
##   2   5  tot_chol  glucose
##   3   1  tot_chol  glucose
##   3   2  tot_chol  glucose
##   3   3  tot_chol  glucose
##   3   4  tot_chol  glucose
##   3   5  tot_chol  glucose
##   4   1  tot_chol  glucose
##   4   2  tot_chol  glucose
##   4   3  tot_chol  glucose
##   4   4  tot_chol  glucose
##   4   5  tot_chol  glucose
##   5   1  tot_chol  glucose
##   5   2  tot_chol  glucose
##   5   3  tot_chol  glucose
##   5   4  tot_chol  glucose
##   5   5  tot_chol  glucose
```



Red: imputed tot_chol vs. Blue: observed tot_chol



Red: the imputed vs. Blue: the observed