Pembersihan Data

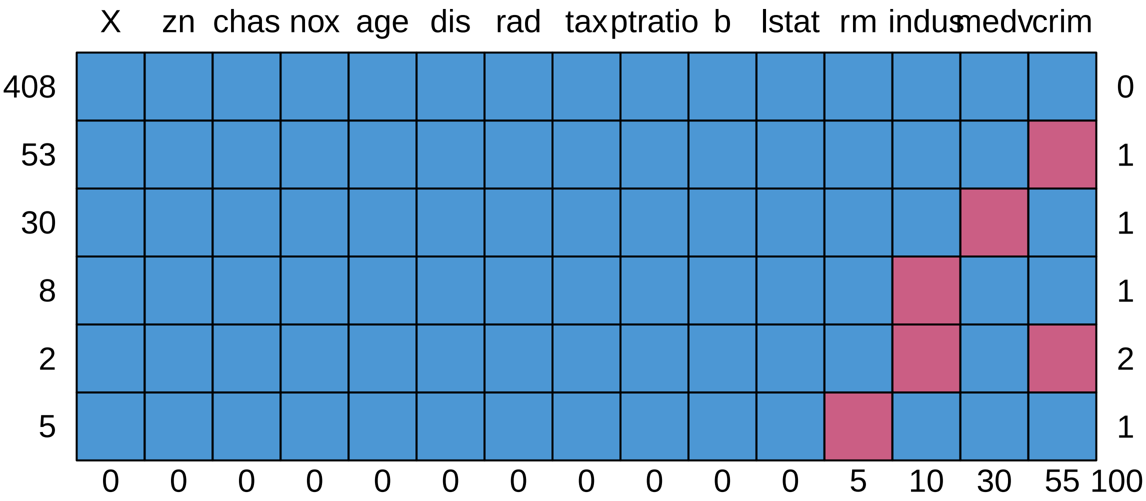
Pembersihan Data

1. Kenalpasti corak data-data lenyap library(mice)

```
MData = read.csv("G:/My Drive/Master-Data-Science/Semester_1/Data_Mining/Data/MData.csv", sep = ";")
head(MData)
```

```
##   X    crim zn indus chas   nox    rm  age    dis rad tax ptratio      b lstat
## 1 1      NA 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98
## 2 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14
## 3 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.03
## 4 4 0.03237  0    NA    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.94
## 5 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90  5.33
## 6 6      NA  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12  5.21
##   medv
## 1 24.0
## 2 21.6
## 3 34.7
## 4 33.4
## 5 36.2
## 6 28.7
```

```
md.pattern(MData)
```



```
##     X zn chas nox age dis rad tax ptratio b lstat rm indus medv crim
## 408 1  1    1   1   1   1   1   1       1 1     1  1     1    1    1   0
## 53  1  1    1   1   1   1   1   1       1 1     1  1     1    0    1
## 30  1  1    1   1   1   1   1   1       1 1     1  0     1    1    1
## 8   1  1    1   1   1   1   1   1       1 1     0  1     1    1    1
## 2   1  1    1   1   1   1   1   1       1 1     0  1     0    2
## 5   1  1    1   1   1   1   1   1       1 1     1  0     1    1    1    1
##     0  0    0   0   0   0   0   0       0 0     0  5    10   30   55  100
```

2. Keluarkan cerapan yang mengandungi data lenyap MData2

```
MData2 = MData[complete.cases(MData),]
head(MData2)
```

```
##     X    crim   zn indus chas   nox    rm  age    dis rad tax ptratio      b
## 2   2 0.02731  0.0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90
## 3   3 0.02729  0.0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83
## 5   5 0.06905  0.0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90
## 7   7 0.08829 12.5  7.87    0 0.524 6.012 66.6 5.5605   5 311    15.2 395.60
## 11 11 0.22489 12.5  7.87    0 0.524 6.377 94.3 6.3467   5 311    15.2 392.52
## 12 12 0.11747 12.5  7.87    0 0.524 6.009 82.9 6.2267   5 311    15.2 396.90
##    lstat medv
## 2   9.14 21.6
## 3   4.03 34.7
## 5   5.33 36.2
## 7  12.43 22.9
## 11 20.45 15.0
## 12 13.27 18.9
```

2.1 Lihat cerapan yang mempunyai data lenyap

```
MData[!complete.cases(MData),]
```

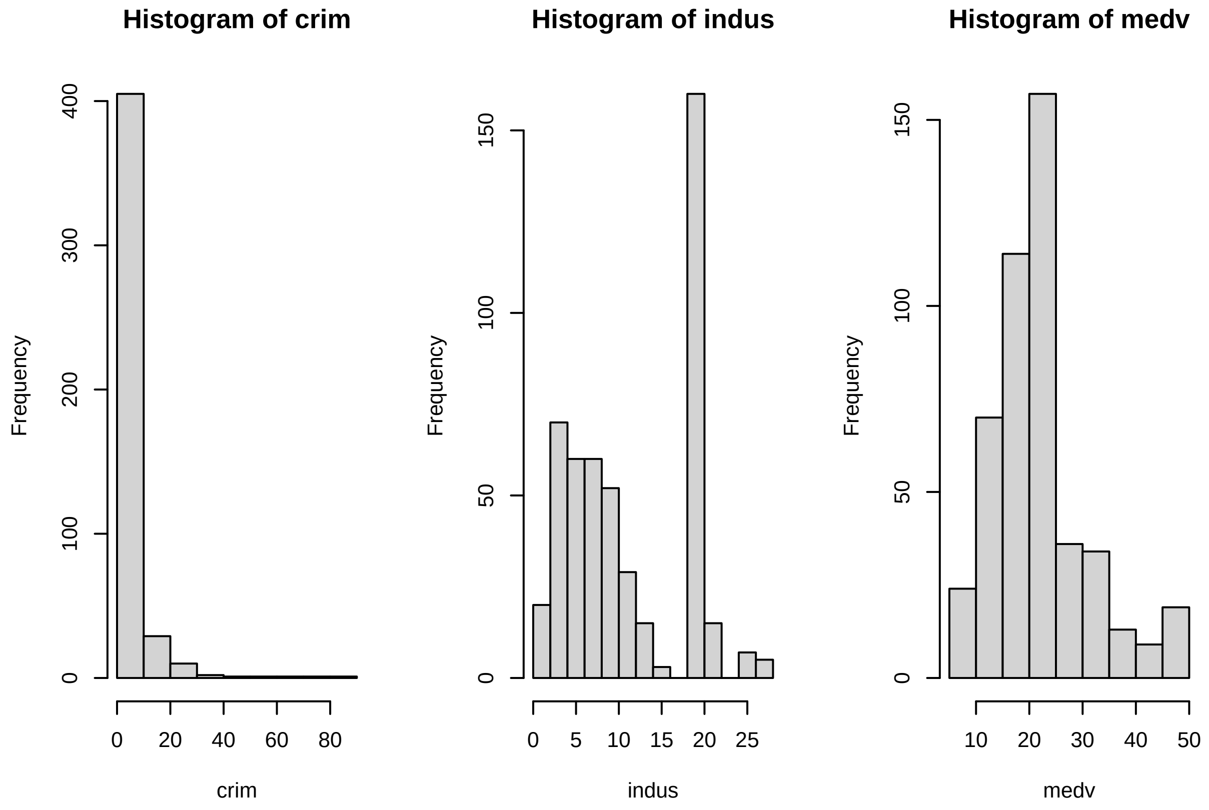
```
##       X    crim   zn indus chas    nox    rm   age    dis rad tax ptratio
## 1     1      NA 18.0  2.31    0 0.5380 6.575  65.2 4.0900   1 296    15.3
## 4     4 0.03237  0.0    NA    0 0.4580 6.998  45.8 6.0622   3 222    18.7
## 6     6      NA  0.0  2.18    0 0.4580 6.430  58.7 6.0622   3 222    18.7
## 8     8 0.14455 12.5  7.87    0 0.5240 6.172  96.1 5.9505   5 311    15.2
## 9     9      NA 12.5  7.87    0 0.5240 5.631 100.0 6.0821   5 311    15.2
## 10   10 0.17004 12.5  7.87    0 0.5240 6.004  85.9 6.5921   5 311    15.2
## 14   14      NA  0.0  8.14    0 0.5380 5.949  61.8 4.7075   4 307    21.0
## 22   22      NA  0.0  8.14    0 0.5380 5.965  89.2 4.0123   4 307    21.0
## 24   24 0.98843  0.0  8.14    0 0.5380 5.813 100.0 4.0952   4 307    21.0
## 36   36      NA  0.0  5.96    0 0.4990 5.933  68.2 3.3603   5 279    19.2
## 41   41 0.03359 75.0  2.95    0 0.4280 7.024  15.8 5.4011   3 252    18.3
## 47   47 0.18836  0.0  6.91    0 0.4480 5.786  33.3 5.1004   3 233    17.9
## 48   48 0.22927  0.0  6.91    0 0.4480    NA  85.5 5.6894   3 233    17.9
## 55   55      NA 75.0  4.00    0 0.4100 5.888  47.6 7.3197   3 469    21.1
## 56   56 0.01311 90.0  1.22    0 0.4030 7.249  21.9 8.6966   5 226    17.9
## 59   59 0.15445 25.0  5.13    0 0.4530 6.145  29.2 7.8148   8 284    19.7
## 61   61      NA 25.0  5.13    0 0.4530 5.741  66.2 7.2254   8 284    19.7
## 66   66 0.03584 80.0    NA    0 0.3980 6.290  17.8 6.6115   4 337    16.1
## 70   70      NA 12.5  6.07    0 0.4090 5.885  33.0 6.4980   4 345    18.9
## 82   82      NA 25.0  4.86    0 0.4260 6.619  70.4 5.4007   4 281    19.0
## 84   84      NA 25.0  4.86    0 0.4260 6.167  46.7 5.4007   4 281    19.0
## 87   87 0.05188  0.0  4.49    0 0.4490 6.015  45.1 4.4272   3 247    18.5
## 88   88 0.07151  0.0  4.49    0 0.4490 6.121  56.8 3.7476   3 247    18.5
## 94   94      NA 28.0 15.04    0 0.4640 6.211  28.9 3.6659   4 270    18.2
## 96   96      NA  0.0  2.89    0 0.4450 6.625  57.8 3.4952   2 276    18.0
## 102 102 0.11432  0.0  8.56    0 0.5200 6.781  71.3 2.8561   5 384    20.9
## 103 103      NA  0.0  8.56    0 0.5200 6.405  85.4 2.7147   5 384    20.9
## 104 104      NA  0.0  8.56    0 0.5200 6.137  87.4 2.7147   5 384    20.9
## 106 106      NA  0.0  8.56    0 0.5200 5.851  96.7 2.1069   5 384    20.9
## 107 107 0.17120  0.0    NA    0 0.5200 5.836  91.9 2.2110   5 384    20.9
## 120 120 0.14476  0.0 10.01    0 0.5470 5.731  65.2 2.7592   6 432    17.8
## 124 124      NA  0.0 25.65    0 0.5810 5.856  97.0 1.9444   2 188    19.1
## 129 129      NA  0.0 21.89    0 0.6240 6.431  98.8 1.8125   4 437    21.2
## 130 130      NA  0.0 21.89    0 0.6240 5.637  94.7 1.9799   4 437    21.2
## 133 133      NA  0.0 21.89    0 0.6240 6.372  97.9 2.3274   4 437    21.2
## 135 135      NA  0.0 21.89    0 0.6240 5.757  98.4 2.3460   4 437    21.2
## 150 150      NA  0.0 19.58    0 0.8710 5.597  94.9 1.5257   5 403    14.7
## 152 152      NA  0.0 19.58    0 0.8710 5.404 100.0 1.5916   5 403    14.7
## 157 157      NA  0.0 19.58    0 0.8710 5.272  94.0 1.7364   5 403    14.7
## 159 159 1.34284  0.0 19.58    0 0.6050    NA 100.0 1.7573   5 403    14.7
## 160 160      NA  0.0 19.58    0 0.8710 6.510 100.0 1.7659   5 403    14.7
## 161 161  1.27346  0.0 19.58    1 0.6050 6.250  92.6  1.7984   5 403    14.7
## 163 163       NA  0.0 19.58    1 0.6050 7.802  98.2  2.0407   5 403    14.7
## 164 164  1.51902  0.0 19.58    1 0.6050 8.375  93.9  2.1620   5 403    14.7
## 171 171       NA  0.0 19.58    0 0.6050 5.875  94.6  2.4259   5 403    14.7
## 174 174       NA  0.0  4.05    0 0.5100 6.416  84.1  2.6463   5 296    16.6
## 181 181       NA  0.0  2.46    0 0.4880 7.765  83.3  2.7410   3 193    17.8
## 182 182  0.06888  0.0  2.46    0 0.4880 6.144  62.2  2.5979   3 193    17.8
## 187 187  0.05602  0.0    NA    0 0.4880 7.831  53.6  3.1992   3 193    17.8
## 196 196  0.01381 80.0  0.46    0 0.4220 7.875  32.0  5.6484   4 255    14.4
## 201 201       NA 95.0  1.47    0 0.4030 7.135  13.9  7.6534   3 402    17.0
## 207 207       NA  0.0 10.59    0 0.4890 6.326  52.5  4.3549   4 277    18.6
## 224 224  0.61470  0.0    NA    0 0.5070 6.618  80.8  3.2721   8 307    17.4
## 225 225       NA  0.0  6.20    0 0.5040 8.266  78.3  2.8944   8 307    17.4
## 243 243  0.10290 30.0    NA    0 0.4280 6.358  52.9  7.0355   6 300    16.6
## 250 250  0.19073 22.0  5.86    0 0.4310 6.718  17.5  7.8265   7 330    19.1
## 258 258       NA 20.0  3.97    0 0.6470 8.704  86.9  1.8010   5 264    13.0
## 266 266  0.76162 20.0  3.97    0 0.6470 5.560  62.8  1.9865   5 264    13.0
## 284 284  0.01501 90.0  1.21    1 0.4010 7.923  24.8  5.8850   1 198    13.6
## 289 289  0.04590 52.5    NA    0 0.4050 6.315  45.6  7.3172   6 293    16.6
## 292 292  0.07886 80.0  4.95    0 0.4110 7.148  27.7  5.1167   4 245    19.2
## 295 295       NA  0.0 13.92    0 0.4370 6.009  42.3  5.5027   4 289    16.0
## 305 305  0.05515 33.0  2.18    0 0.4720 7.236  41.1  4.0220   7 222    18.4
## 310 310       NA  0.0  9.90    0 0.5440 5.972  76.7  3.1025   4 304    18.4
## 311 311       NA  0.0  9.90    0 0.5440 4.973  37.8  2.5194   4 304    18.4
## 319 319  0.40202  0.0  9.90    0 0.5440 6.382  67.2  3.5325   4 304    18.4
## 331 331  0.04544  0.0  3.24    0 0.4600 6.144  32.2  5.8736   4 430    16.9
## 333 333       NA 35.0    NA    0 0.4379 6.031  23.3  6.6407   1 304    16.9
## 334 334       NA  0.0  5.19    0 0.5150 6.316  38.1  6.4584   5 224    20.2
## 336 336  0.03961  0.0  5.19    0 0.5150 6.037  34.5  5.9853   5 224    20.2
## 347 347       NA  0.0  4.39    0 0.4420 5.898  52.3  8.0136   3 352    18.8
## 349 349       NA 80.0  2.01    0 0.4350 6.635  29.7  8.3440   4 280    17.0
## 354 354  0.01709 90.0  2.02    0 0.4100 6.728  36.1 12.1265   5 187    17.0
## 364 364       NA  0.0 18.10    1 0.7700 5.803  89.0  1.9047  24 666    20.2
## 365 365       NA  0.0 18.10    1 0.7180 8.780  82.9  1.9047  24 666    20.2
## 366 366       NA  0.0 18.10    0 0.7180 3.561  87.9  1.6132  24 666    20.2
## 368 368 13.52220  0.0 18.10    0 0.6310 3.863 100.0  1.5106  24 666    20.2
## 372 372  9.23230  0.0 18.10    0 0.6310    NA 100.0  1.1691  24 666    20.2
## 383 383       NA  0.0 18.10    0 0.7000 5.536 100.0  1.5804  24 666    20.2
## 400 400  9.91655  0.0 18.10    0 0.6930    NA  77.8  1.5004  24 666    20.2
## 402 402       NA  0.0 18.10    0 0.6930 6.343 100.0  1.5741  24 666    20.2
## 413 413 18.81100  0.0    NA    0 0.5970 4.628 100.0  1.5539  24 666    20.2
## 415 415       NA  0.0 18.10    0 0.6930 4.519 100.0  1.6582  24 666    20.2
## 418 418 25.94060  0.0 18.10    0 0.6790 5.304  89.1  1.6475  24 666    20.2
## 421 421       NA  0.0 18.10    0 0.7180 6.411 100.0  1.8589  24 666    20.2
## 423 423       NA  0.0 18.10    0 0.6140 5.648  87.6  1.9512  24 666    20.2
## 425 425  8.79212  0.0 18.10    0 0.5840    NA  70.6  2.0635  24 666    20.2
## 435 435       NA  0.0    NA    0 0.7130 6.208  95.0  2.2222  24 666    20.2
## 437 437       NA  0.0 18.10    0 0.7400 6.461  93.3  2.0026  24 666    20.2
## 438 438       NA  0.0 18.10    0 0.7400 6.152 100.0  1.9142  24 666    20.2
## 445 445       NA  0.0 18.10    0 0.7400 5.854  96.6  1.8956  24 666    20.2
## 453 453       NA  0.0 18.10    0 0.7130 6.297  91.8  2.3682  24 666    20.2
## 476 476  6.39312  0.0 18.10    0 0.5840 6.162  97.4  2.2060  24 666    20.2
## 492 492       NA  0.0 27.74    0 0.6090 5.983  98.8  1.8681   4 711    20.1
## 496 496  0.17899  0.0  9.69    0 0.5850 5.670  28.8  2.7986   6 391    19.2
## 497 497       NA  0.0  9.69    0 0.5850 5.390  72.9  2.7986   6 391    19.2
## 499 499  0.23912  0.0  9.69    0 0.5850 6.019  65.3  2.4091   6 391    19.2
## 503 503  0.04527  0.0 11.93    0 0.5730 6.120  76.7  2.2875   1 273    21.0
##          b lstat medv
## 1   396.90  4.98 24.0
## 4   394.63  2.94 33.4
## 6   394.12  5.21 28.7
## 8   396.90 19.15   NA
## 9   386.63 29.93 16.5
## 10  386.71 17.10   NA
## 14  396.90  8.26 20.4
## 22  392.53 13.83 19.6
## 24  394.54 19.88   NA
## 36  396.90  9.68 18.9
## 41  395.62  1.98   NA
## 47  396.90 14.15   NA
## 48  392.74 18.80 16.6
## 55  396.90 14.80 18.9
## 56  395.93  4.81   NA
## 59  390.68  6.86   NA
## 61  395.11 13.15 18.7
## 66  396.90  4.67 23.5
## 70  396.90  8.79 20.9
## 82  395.63  7.22 23.9
## 84  390.64  7.51 22.9
## 87  395.99 12.86   NA
## 88  395.15  8.44   NA
## 94  396.33  6.21 25.0
## 96  357.98  6.65 28.4
## 102 395.58  7.67   NA
## 103  70.80 10.63 18.6
## 104 394.47 13.44 19.3
## 106 394.05 16.47 19.5
## 107 395.67 18.66 19.5
## 120 391.50 13.61   NA
## 124 370.31 25.41 17.3
## 129 396.90 15.39 18.0
## 130 396.90 18.34 14.3
## 133 385.76 11.12 23.0
## 135 262.76 17.31 15.6
## 150 351.85 21.45 15.4
## 152 341.60 13.28 19.6
## 157  88.63 16.14 13.1
## 159 353.89  6.43 24.3
## 160 364.31  7.39 23.3
## 161 338.92  5.50   NA
## 163 389.61  1.92 50.0
## 164 388.45  3.32   NA
## 171 292.29 14.43 17.4
## 174 395.50  9.04 23.6
## 181 395.56  7.56 39.8
## 182 396.90  9.45   NA
## 187 392.63  4.45 50.0
## 196 394.23  2.97   NA
## 201 384.30  4.45 32.9
## 207 394.87 10.97 24.4
## 224 396.90  7.60 30.1
## 225 385.05  4.14 44.8
## 243 372.75 11.22 22.2
## 250 393.74  6.56   NA
## 258 389.70  5.12 50.0
## 266 392.40 10.45   NA
## 284 395.52  3.16   NA
## 289 396.90  7.60 22.3
## 292 396.90  3.56   NA
## 295 396.90 10.40 21.7
## 305 393.68  6.93   NA
## 310 396.24  9.97 20.3
## 311 350.45 12.64 16.1
## 319 395.21 10.36   NA
## 331 368.57  9.09   NA
## 333 362.25  7.83 19.4
## 334 389.71  5.68 22.2
## 336 396.90  8.01   NA
## 347 364.61 12.67 17.2
## 349 390.94  5.99 24.5
## 354 384.46  4.50   NA
## 364 353.04 14.64 16.8
## 365 354.55  5.29 21.9
## 366 354.70  7.12 27.5
## 368 131.42 13.33   NA
## 372 366.15  9.53 50.0
## 383 396.90 23.60 11.3
## 400 338.16 29.97  6.3
## 402 396.90 20.32  7.2
## 413  28.79 34.37 17.9
## 415  88.27 36.98  7.0
## 418 127.36 26.64   NA
## 421 318.75 15.02 16.7
## 423 291.55 14.10 20.8
## 425   3.65 17.16 11.7
## 435 100.63 15.17 11.7
## 437  27.49 18.05  9.6
## 438   9.32 26.45  8.7
## 445 240.52 23.79 10.8
## 453 385.09 17.27 16.1
## 476 302.76 24.10   NA
## 492 390.11 18.07 13.6
## 496 393.29 17.60   NA
## 497 396.90 21.14 19.7
## 499 396.90 12.92   NA
## 503 396.90  9.08   NA
```

3. Lengkapkan data lenyap secara manual

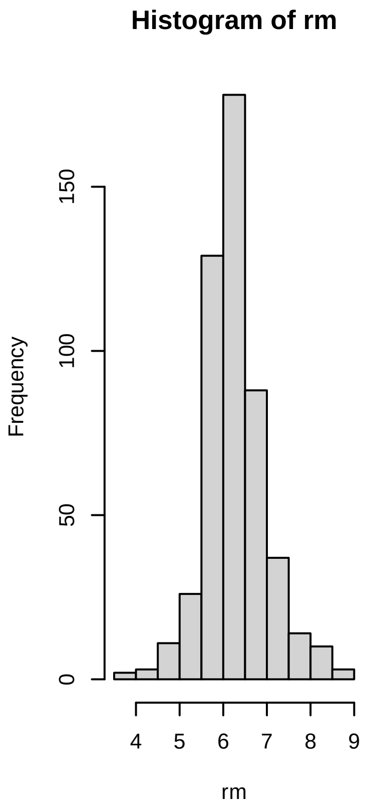
```r
# indus.fix = edit(MData$indus) # Tak tukar ori data
# head(indus.fix)
```

4. Gunakan sukatan memusat sebagai anggaran terhadap data lenyap

```r
attach(MData)

par(mfrow =c(1,3))
hist(crim) #tak simetri
hist(indus) #tak simteri
hist(medv) # tak simetri
```



```r
hist(rm) # simetri
```



4.2 Untuk data taburan bersifat pincang/bukan simetri: median boleh digunakan.

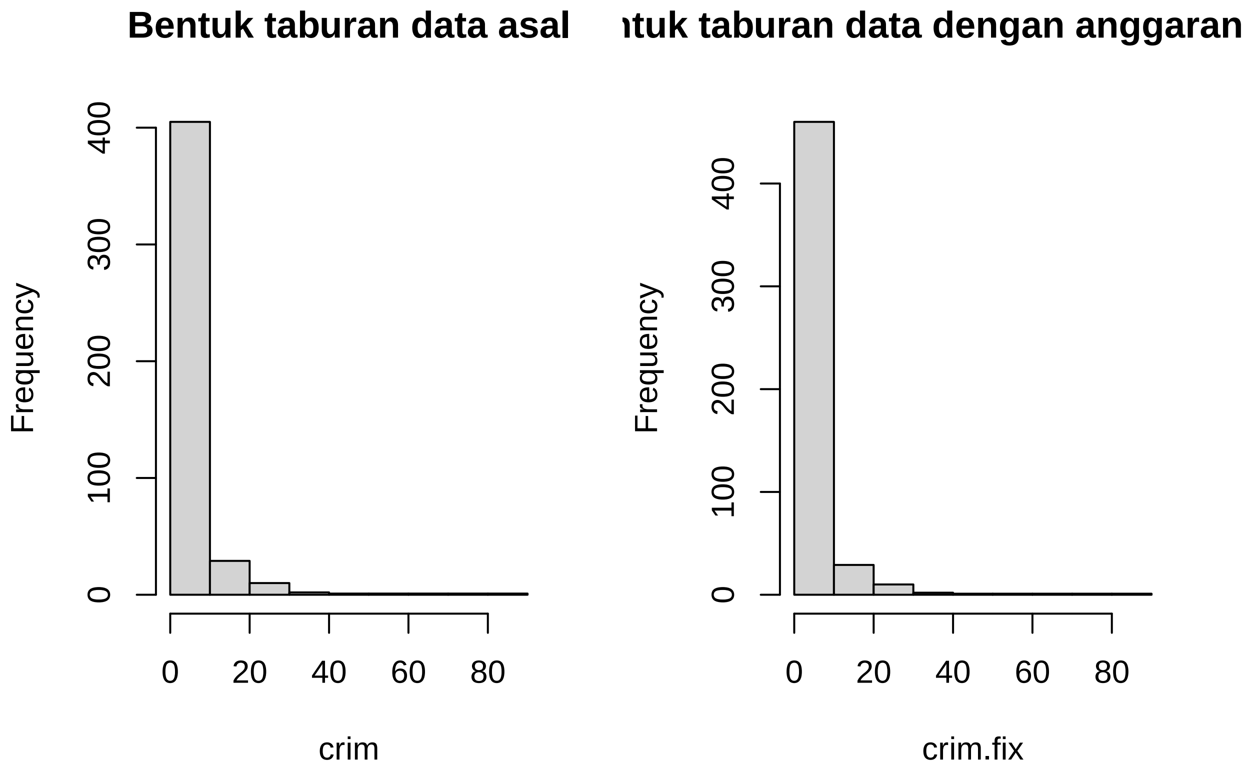
Kenal pasti median data

```
median.crim = median(crim, na.rm=T)
median.crim
```

crim

```
## [1] 0.25199
```

```
crim.fix = ifelse(is.na(crim), median.crim, crim)

par(mfrow = c(1,2))
hist(crim, main="Bentuk taburan data asal")
hist(crim.fix, main="Bentuk taburan data dengan anggaran median")
```

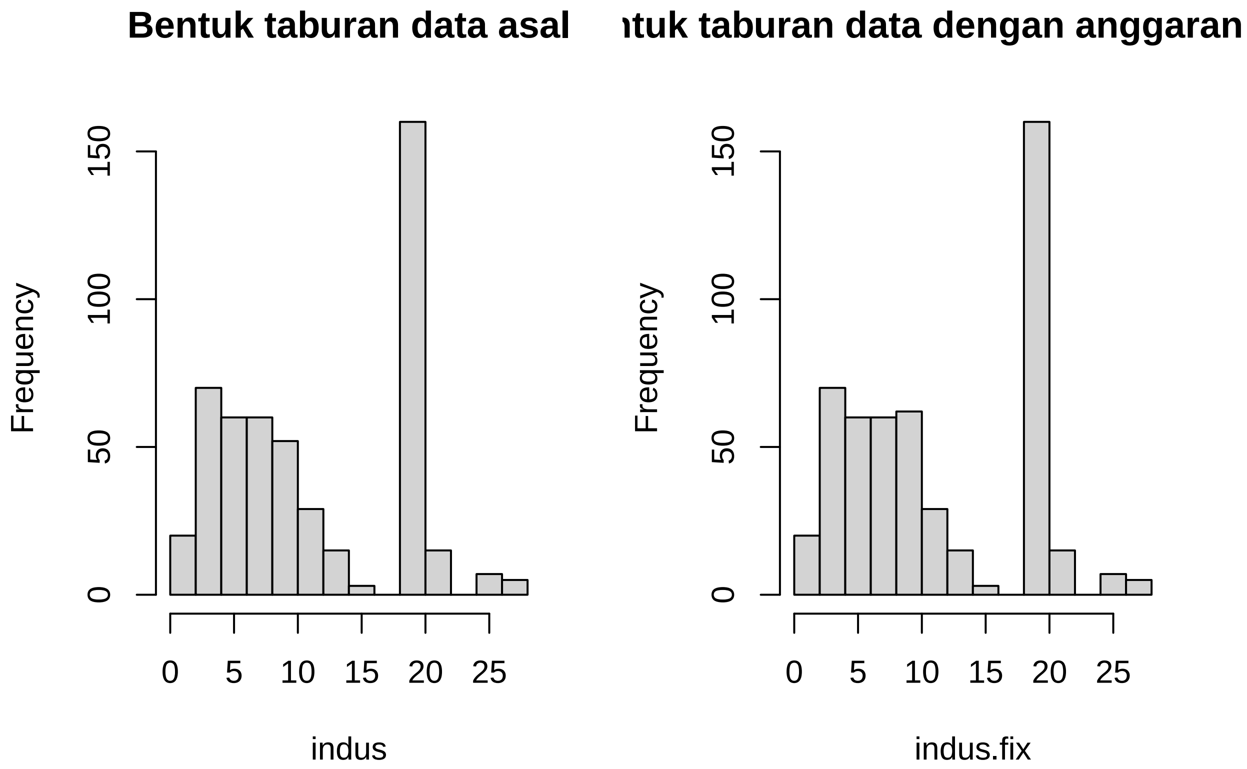


```
median.indus = median(indus, na.rm=T)
median.crim
```

indus

```
## [1] 0.25199
```

```
indus.fix = ifelse(is.na(indus), median.indus, indus)

par(mfrow = c(1,2))
hist(indus, main="Bentuk taburan data asal")
hist(indus.fix, main="Bentuk taburan data dengan anggaran median")
```

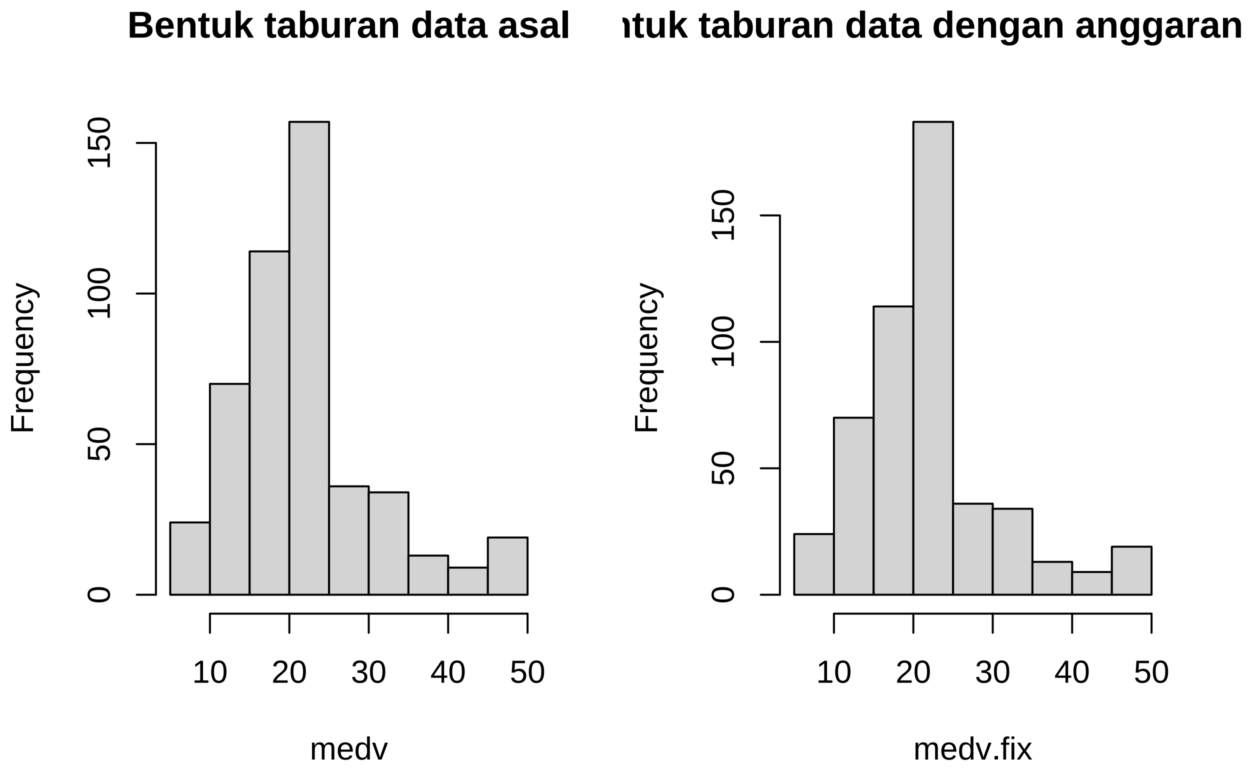


```
median.medv = median(medv, na.rm=T)
median.medv
```

medv

```
## [1] 20.95
```

```
medv.fix = ifelse(is.na(medv), median.medv, medv)

par(mfrow = c(1,2))
hist(medv, main="Bentuk taburan data asal")
hist(medv.fix, main="Bentuk taburan data dengan anggaran median")
```



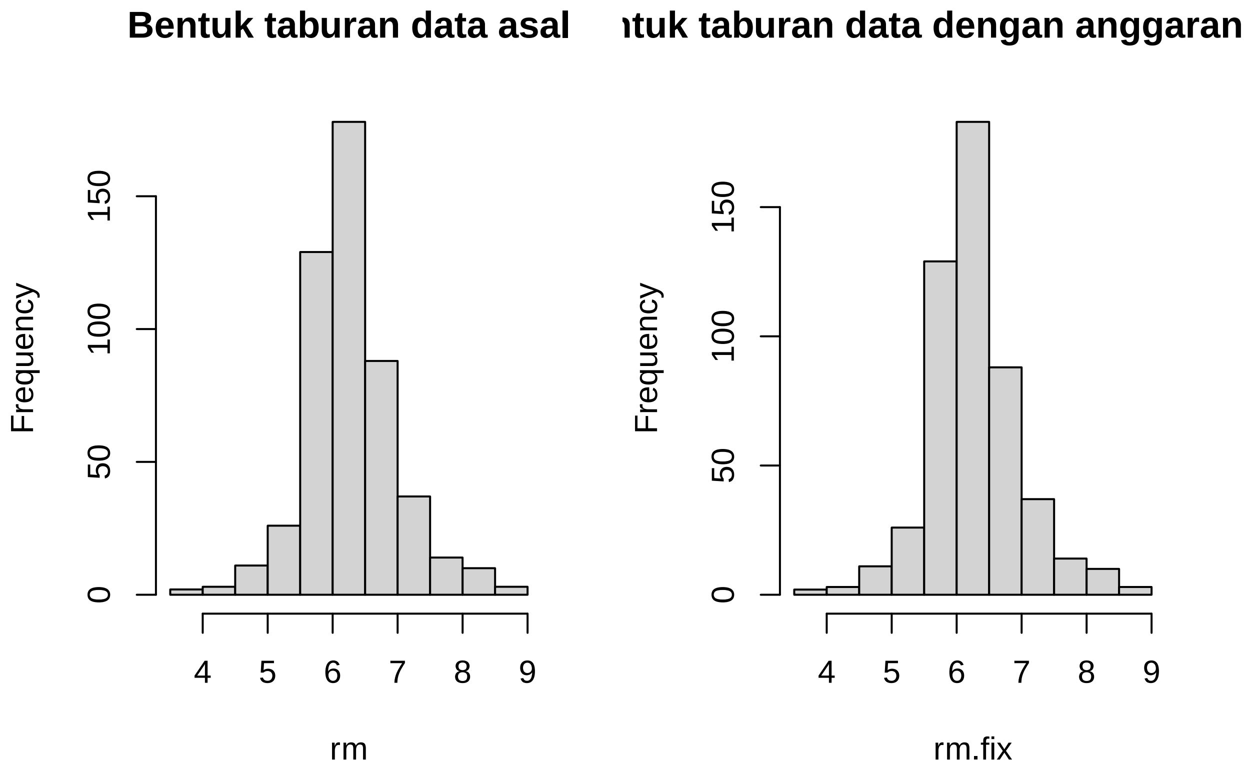
4.3 Untuk data taburan normal/simetri dengan nilai berangka: nilai minboleh digunakan.

```
mean.rm = mean(rm, na.rm=T)
mean.rm
```

rm

```
## [1] 6.288016
```

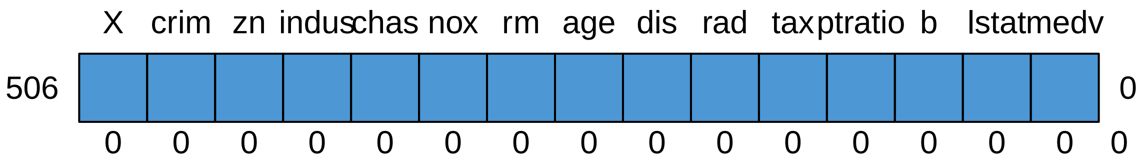
```
rm.fix = ifelse(is.na(rm),mean.rm, rm)

par(mfrow = c(1,2))
hist(rm, main="Bentuk taburan data asal")
hist(rm.fix, main="Bentuk taburan data dengan anggaran median")
```



Bentukkan set data lengkap

```
MData.lengkap = MData
MData.lengkap$crim = crim.fix
MData.lengkap$medv = medv.fix
MData.lengkap$indus = indus.fix
MData.lengkap$rm = rm.fix
md.pattern(MData.lengkap)
```

```
##  /\     /\
## {  `---'  }
## {  O   O  }
## ==>  V <==  No need for mice. This data set is completely observed.
##  \  \|/  /
##   `-----'
```



```
##     X crim zn indus chas nox rm age dis rad tax ptratio b lstat medv
## 506 1    1  1     1    1   1  1   1   1   1   1       1 1     1    1 0
##     0    0  0     0    0   0  0   0   0   0   0       0 0     0    0 0
```

```
MData.lengkap = MData.lengkap[-1]
head(MData.lengkap)
```

```
##      crim zn indus chas   nox    rm  age    dis rad tax ptratio      b lstat
## 1 0.25199 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.03
## 4 0.03237  0  9.69    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.94
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90  5.33
## 6 0.25199  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12  5.21
##   medv
## 1 24.0
## 2 21.6
## 3 34.7
## 4 33.4
## 5 36.2
## 6 28.7
```

5. Gunakan maklumat k-jiran terdekat sebagai anggaran terhadap data lenyap

```
iris.mis1 = read.csv("G:/My Drive/Master-Data-Science/Semester_1/Data_Mining/Data/iris.mis1.csv")
iris.mis1 = iris.mis1[-1]

library(multiUS)

iris.mis1 = KNNimp(data=iris.mis1, k=10)
head(iris.mis1)
```

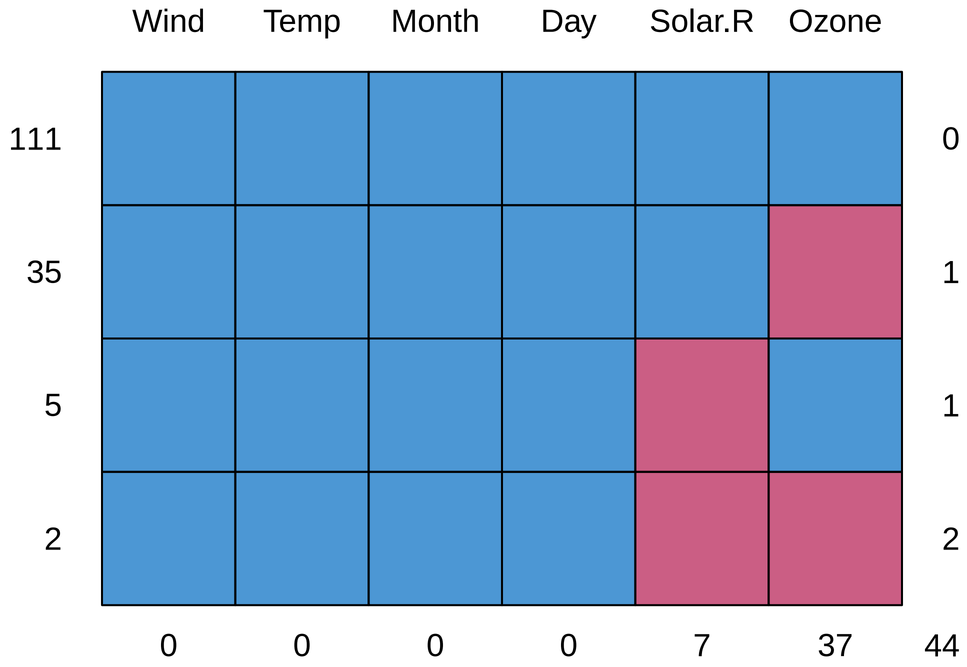
```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1      5.09944    3.500000          1.4         0.2
## 2      4.90000    3.000000          1.4         0.2
## 3      4.70000    3.100142          1.3         0.2
## 4      4.60000    3.100000          1.5         0.2
## 5      5.00000    3.600000          1.4         0.2
## 6      5.40000    3.900000          1.7         0.4
```

6. Anggaran data lenyap menerusi pelbagai kaedah imputasi statistik: (pakej mice)

6.1 Data dengan p/ubah nilai berangka

model = predictive mean matching

```
airquality = read.table("G:/My Drive/Master-Data-Science/Semester_1/Data_Mining/Data/airquality.txt", he
par(mfrow = c(1,1))
md.pattern(airquality)
```

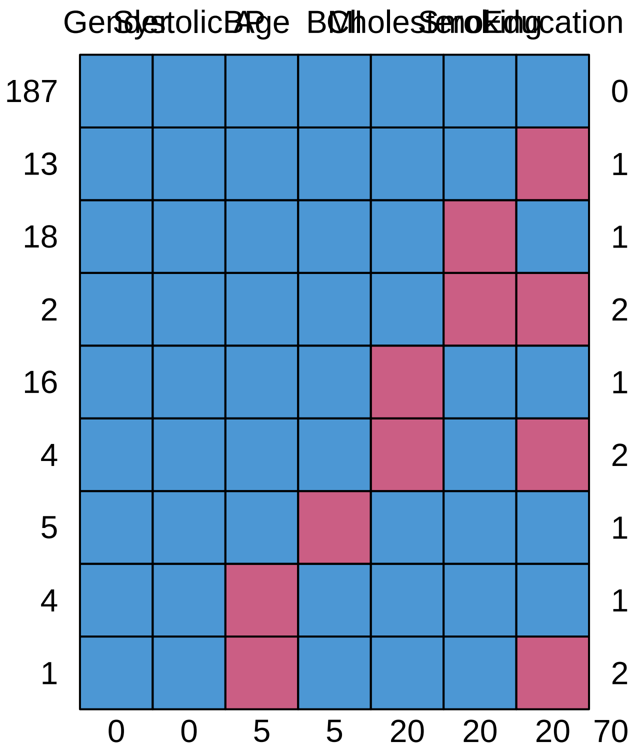


```
##     Wind Temp Month Day Solar.R Ozone
## 111    1    1     1   1       1     1  0
## 35     1    1     1   1       1     0  1
## 5      1    1     1   1       0     1  1
## 2      1    1     1   1       0     0  2
##        0    0     0   0       7    37 44
```

6.2 Data dengan p/ubah nilai berbeza

model = Logistic Regression

```
data2 = read.csv("G:/My Drive/Master-Data-Science/Semester_1/Data_Mining/Data/dat2.csv")
data2 = data2[-1]
md.pattern(data2)
```



```
##     Gender SystolicBP Age BMI Cholesterol Smoking Education
## 187      1          1   1   1           1       1         1  0
## 13       1          1   1   1           1       1         0  1
## 18       1          1   1   1           1       0         1  1
## 2        1          1   1   1           1       0         0  2
## 16       1          1   1   1           0       1         1  1
## 4        1          1   1   1           0       1         0  2
## 5        1          1   1   0           1       1         1  1
## 4        1          1   0   1           1       1         1  1
## 1        1          1   0   1           1       1         0  2
##          0          0   5   5          20      20        20 70
```

```
str(data2)
```

```
## 'data.frame':    250 obs. of  7 variables:
##  $ Age        : num  67.9 54.8 68.4 67.9 60.9 44.9 49.9 55.1 57.5 77.2 ...
##  $ Gender     : chr  "Female" "Female" "Male" "Male" ...
##  $ Cholesterol: num  236 256 199 205 208 ...
##  $ SystolicBP : num  130 133 158 136 145 ...
##  $ BMI        : num  26.4 28.4 24.1 19.9 26.7 30.6 27.3 27.5 28.3 29.1 ...
##  $ Smoking    : chr  "Yes" "No" "Yes" "No" ...
##  $ Education  : chr  "High" "Medium" "High" "Low" ...
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
dat = data2%>%
            mutate(Smoking = as.factor(Smoking)) %>%
            mutate(Education = factor(Education, levels = c("Low","Medium","High"), ordered=T)) %>%
            mutate(Gender = as.factor(Gender))
```

Imputasi data

```
init = mice(dat, maxit=0)
meth = init$method
predM = init$predictorMatrix
```

setkan kaedah imputasi yang digunakan

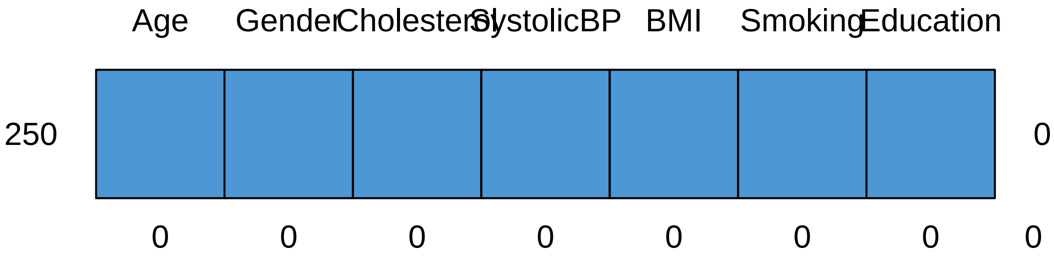
Setiap p/ubah akan mengambil kaedah yang berbeza mengikut jenis data

```
meth[c('Age')] = "pmm"
meth[c('Cholesterol')] = "pmm"
meth[c('SystolicBP')] = "pmm"
meth[c('BMI')] = "pmm"
meth[c('Gender')] = "logreg"
meth[c('Smoking')] = "logreg"
meth[c('Education')] = "polyreg"

ImputedData = mice(dat, method=meth, predictorMatrix = predM)
```

```
##
##  iter imp variable
##   1   1  Age  Cholesterol  BMI  Smoking  Education
##   1   2  Age  Cholesterol  BMI  Smoking  Education
##   1   3  Age  Cholesterol  BMI  Smoking  Education
##   1   4  Age  Cholesterol  BMI  Smoking  Education
##   1   5  Age  Cholesterol  BMI  Smoking  Education
##   2   1  Age  Cholesterol  BMI  Smoking  Education
##   2   2  Age  Cholesterol  BMI  Smoking  Education
##   2   3  Age  Cholesterol  BMI  Smoking  Education
##   2   4  Age  Cholesterol  BMI  Smoking  Education
##   2   5  Age  Cholesterol  BMI  Smoking  Education
##   3   1  Age  Cholesterol  BMI  Smoking  Education
##   3   2  Age  Cholesterol  BMI  Smoking  Education
##   3   3  Age  Cholesterol  BMI  Smoking  Education
##   3   4  Age  Cholesterol  BMI  Smoking  Education
##   3   5  Age  Cholesterol  BMI  Smoking  Education
##   4   1  Age  Cholesterol  BMI  Smoking  Education
##   4   2  Age  Cholesterol  BMI  Smoking  Education
##   4   3  Age  Cholesterol  BMI  Smoking  Education
##   4   4  Age  Cholesterol  BMI  Smoking  Education
##   4   5  Age  Cholesterol  BMI  Smoking  Education
##   5   1  Age  Cholesterol  BMI  Smoking  Education
##   5   2  Age  Cholesterol  BMI  Smoking  Education
##   5   3  Age  Cholesterol  BMI  Smoking  Education
##   5   4  Age  Cholesterol  BMI  Smoking  Education
##   5   5  Age  Cholesterol  BMI  Smoking  Education
```

```
CompletedData = complete(ImputedData)
md.pattern(CompletedData)
```

```
##  /\     /\
## {  `---'  }
## {  O   O  }
## ==>  V <==  No need for mice. This data set is completely observed.
##  \  \|/  /
##   `-----'
```



```
##     Age Gender Cholesterol SystolicBP BMI Smoking Education
## 250   1      1           1          1   1       1         1 0
##       0      0           0          0   0       0         0 0
```

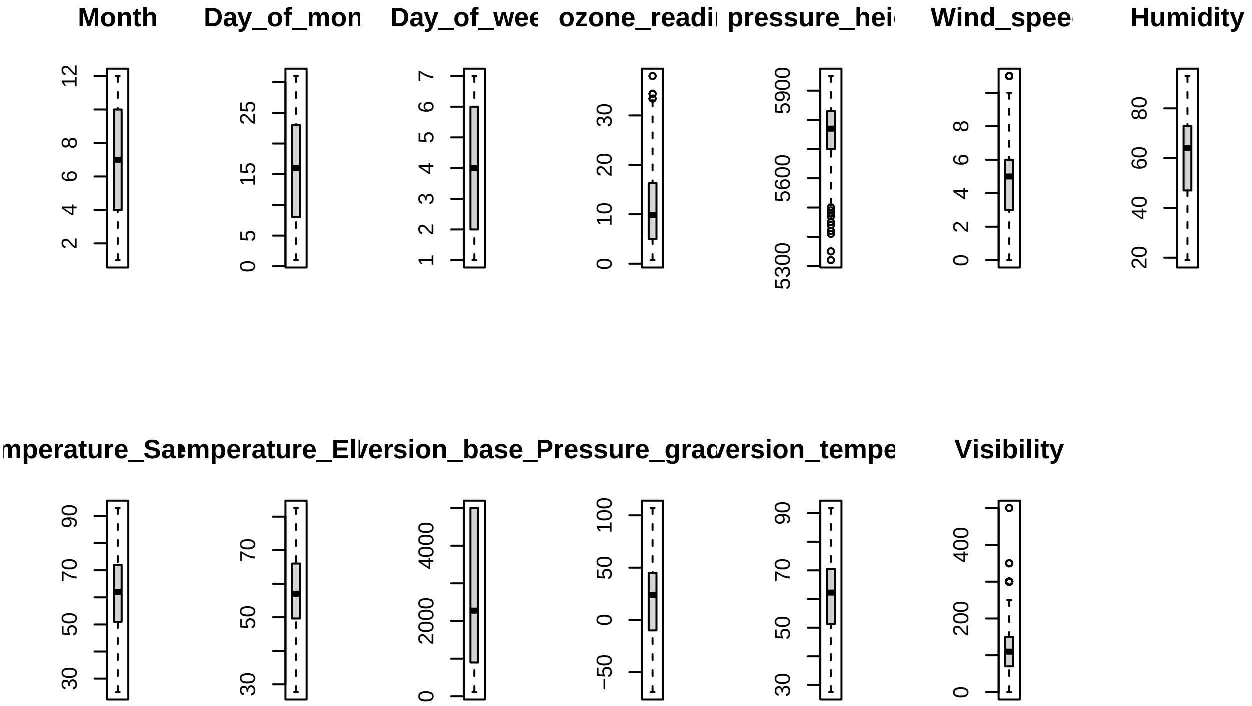
Mengurus Data Pencil

1. Pendekatan Univariat (satu p/ubah)

```
ozone3 = read.csv("G:/My Drive/Master-Data-Science/Semester_1/Data_Mining/Data/ozone3.csv", header=T)
ozone3 = ozone3[,-1]
attach(ozone3)
str(ozone3)
```

```
## ’data.frame':    366 obs. of  13 variables:
##  $ Month                 : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Day_of_month          : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Day_of_week           : int  4 5 6 7 1 2 3 4 5 6 ...
##  $ ozone_reading         : num  3.01 3.2 2.7 5.18 5.34 5.77 3.69 3.89 5.76 6.94 ...
##  $ pressure_height       : int  5480 5660 5710 5700 5760 5720 5790 5790 5700 5700 ...
##  $ Wind_speed            : int  8 6 4 3 3 4 6 3 3 3 ...
##  $ Humidity              : int  20 32 28 37 51 69 19 25 73 59 ...
##  $ Temperature_Sandburg : int  30 38 40 45 54 35 45 55 41 44 ...
##  $ Temperature_ElMonte  : num  32.5 41.4 38.1 47.1 45.3 ...
##  $ Inversion_base_height: int  5000 1601 2693 590 1450 1568 2631 554 2083 2654 ...
##  $ Pressure_gradient    : int  -15 -14 -25 -24 25 15 -33 -28 23 -2 ...
##  $ Inversion_temperature: num  30.6 46.9 47.7 55 57 ...
##  $ Visibility           : int  200 300 250 100 60 60 100 250 120 120 ...
```

```
par(mfrow = c(2,7))
for (col in names(ozone3)) {
  boxplot(ozone3[[col]], main = col)
}
```



1.1 Mengesan data pencil

Ozone Reading

```
outlier_ozone = boxplot.stats(ozone_reading)$out
out_ind = which(ozone_reading%in%outlier_ozone)

ozone3[c(out_ind),]
```

```
##     Month Day_of_month Day_of_week ozone_reading pressure_height Wind_speed
## 188     7            6           2         34.39            5900          6
## 189     7            7           3         33.40            5890          5
## 243     8           30           1         37.98            5950          5
##     Humidity Temperature_Sandburg Temperature_ElMonte Inversion_base_height
## 188       86                   87               81.68                   990
## 189       65                   91               81.68                   508
## 243       62                   92               82.40                   557
##     Pressure_gradient Inversion_temperature Visibility
## 188                22                 85.10         40
## 189                29                 85.28        100
## 243                 0                 90.68         70
```

Pressure height

```
outlier_PH = boxplot.stats(pressure_height)$out
out_ind = which(pressure_height%in%c(outlier_PH))

ozone3[c(out_ind),]
```

```
##     Month Day_of_month Day_of_week ozone_reading pressure_height Wind_speed
## 1       1            1           4          3.01            5480          8
## 36      2            5           4          2.94            5410          6
## 37      2            6           5          2.74            5350          7
## 38      2            7           6          2.21            5480          9
## 40      2            9           1          2.92            5490         11
## 62      3            2           2          3.22            5470          7
## 63      3            3           3          2.79            5320         11
## 64      3            4           4          5.20            5420          8
## 95      4            4           7          3.82            5420          7
## 104     4           13           2          3.65            5440          5
## 105     4           14           3          6.76            5480          7
## 107     4           16           5          4.34            5450         11
## 317    11           12           5          2.90            5500          9
##     Humidity Temperature_Sandburg Temperature_ElMonte Inversion_base_height
## 1         20                   30               32.54                  5000
## 36        64                   31               32.18                  5000
## 37        62                   30               32.54                  1341
## 38        72                   36               37.58                  5000
## 40        72                   37               38.48                  5000
## 62        46                   30               29.66                  5000
## 63        45                   25               27.68                  5000
## 64        33                   39               30.20                  5000
## 95        69                   35               33.08                  5000
## 104       44                   35               33.08                  5000
## 105       51                   46               37.40                  2490
## 107       35                   32               33.26                  5000
## 317       56                   39               41.36                  5000
##     Pressure_gradient Inversion_temperature Visibility
## 1                 -15                 30.56        200
## 36                 28                 32.36        200
## 37                 18                 45.86         60
## 38                  0                 38.66        350
## 40                 32                 38.12        350
## 62                 44                 29.30        300
## 63                 39                 27.50        200
## 64                 15                 30.02        500
## 95                 41                 30.92        200
## 104                24                 32.54         80
## 105                29                 47.48        300
## 107                36                 33.44        300
## 317                15                 41.72        120
```

Wind Speed

```
outlier_WS = boxplot.stats(Wind_speed)$out
out_ind = which(Wind_speed%in%outlier_WS)
```

```r
ozone3[c(out_ind),]
```

```
##     Month Day_of_month Day_of_week ozone_reading pressure_height Wind_speed
## 40      2            9           1          2.92            5490         11
## 53      2           22           7          3.61            5730         11
## 63      3            3           3          2.79            5320         11
## 107     4           16           5          4.34            5450         11
##     Humidity Temperature_Sandburg Temperature_ElMonte Inversion_base_height
## 40        72                   37               38.48                  5000
## 53        19                   51               55.40                  5000
## 63        45                   25               27.68                  5000
## 107       35                   32               33.26                  5000
##     Pressure_gradient Inversion_temperature Visibility
## 40                 32                 38.12        350
## 53                -43                 49.10        300
## 63                 39                 27.50        200
## 107                36                 33.44        300
```

Visibility

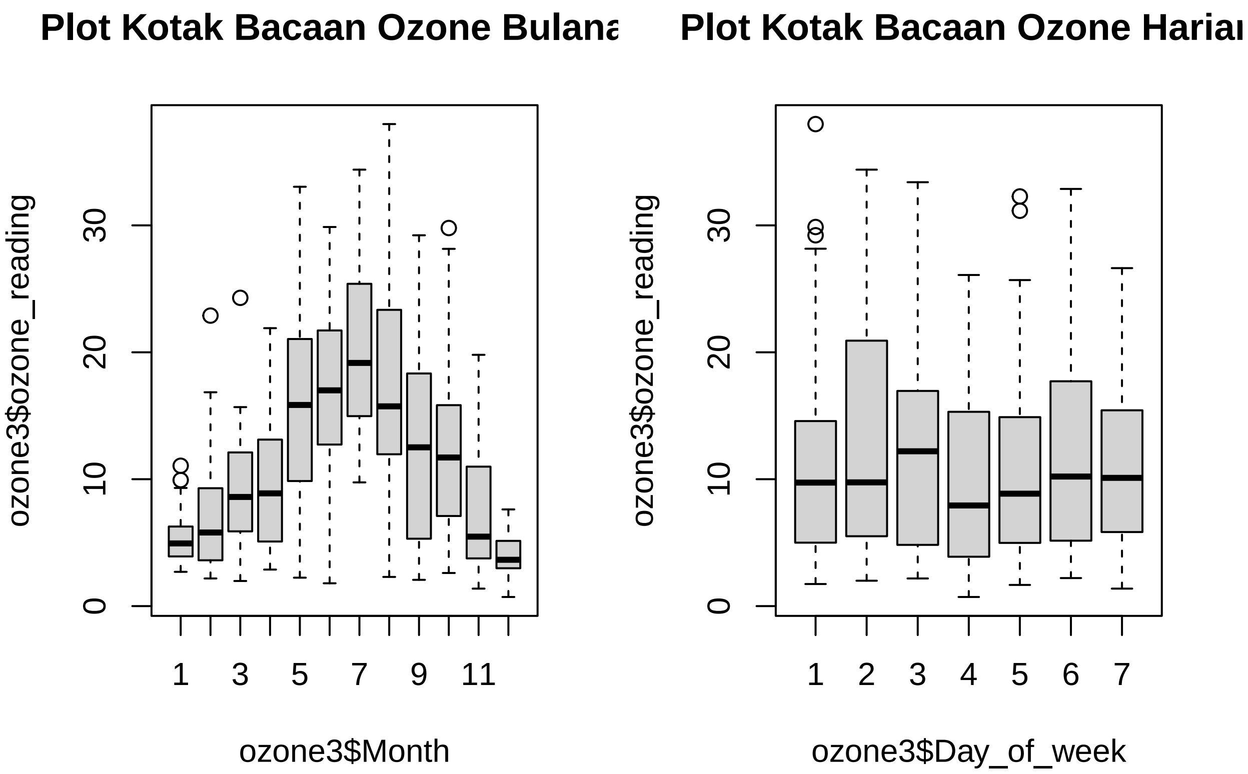
```r
outlier_vis = boxplot.stats(Visibility)$out
out_ind = which(Visibility%in%outlier_vis)

ozone3[c(out_ind),]
```

```
##     Month Day_of_month Day_of_week ozone_reading pressure_height Wind_speed
## 2       1            2           5          3.20            5660          6
## 38      2            7           6          2.21            5480          9
## 40      2            9           1          2.92            5490         11
## 41      2           10           2          4.08            5560         10
## 42      2           11           3          6.04            5700          3
## 43      2           12           4          8.32            5680          5
## 51      2           20           5          5.73            5690          8
## 52      2           21           6          4.85            5700          3
## 53      2           22           7          3.61            5730         11
## 54      2           23           1          4.04            5690          7
## 55      2           24           2          6.04            5640          5
## 62      3            2           2          3.22            5470          7
## 64      3            4           4          5.20            5420          8
## 72      3           12           5          7.63            5690          0
## 73      3           13           6         12.22            5760          4
## 81      3           21           7          8.07            5720          5
## 91      3           31           3         12.33            5710          3
## 97      4            6           2          9.32            5590          6
## 98      4            7           3         13.12            5690          6
## 105     4           14           3          6.76            5480          7
## 107     4           16           5          4.34            5450         11
## 232     8           19           4          8.97            5730          7
## 234     8           21           6         17.18            5790          4
## 236     8           23           1         20.24            5880          3
## 301    10           27           3          2.61            5760          5
## 310    11            5           5          4.91            5860          7
## 318    11           13            6          5.32            5660          3
## 341    12            6            1          4.65            5780          4
## 343    12            8            3          4.31            5760          0
## 357    12           22            3          4.25            5710          4
##     Humidity Temperature_Sandburg Temperature_ElMonte Inversion_base_height
## 2         32                   38               41.36                  1601
## 38        72                   36               37.58                  5000
## 40        72                   37               38.48                  5000
## 41        72                   41               40.46                  5000
## 42        32                   46               48.38                  5000
## 43        50                   51               47.12                  5000
## 51        21                   41               43.88                  5000
## 52        19                   45               48.02                  5000
## 53        19                   51               55.40                  5000
## 54        19                   53               50.18                  5000
## 55        68                   50               37.40                  5000
## 62        46                   30               29.66                  5000
## 64        33                   39               30.20                  5000
## 72        60                   49               46.04                   613
## 73        31                   56               51.80                   334
## 81        19                   59               59.72                   377
## 91        46                   62               52.52                   472
## 97        51                   48               38.12                  5000
## 98        63                   59               52.88                  2014
## 105       51                   46               37.40                  2490
## 107       35                   32               33.26                  5000
## 232       72                   67               57.20                  5000
## 234       57                   74               64.40                   994
## 236       73                   77               66.38                   636
## 301       23                   57               53.42                  5000
## 310       19                   70               62.78                  5000
## 318       54                   50               46.94                  5000
## 341       19                   48               54.14                  2933
## 343       32                   62               56.12                   826
## 357       19                   51               51.08                  5000
##     Pressure_gradient Inversion_temperature Visibility
## 2                 -14                 46.94        300
## 38                  0                 38.66        350
## 40                 32                 38.12        350
## 41                 -1                 37.58        300
## 42                -30                 45.86        300
## 43                 -8                 45.50        300
## 51                -30                 42.26        300
## 52                -53                 43.88        300
## 53                -43                 49.10        300
## 54                  7                 49.10        300
## 55                 24                 42.08        300
## 62                 44                 29.30        300
## 64                 15                 30.02        500
## 72                -27                 59.72        300
## 73                 -9                 64.40        300
## 81                -27                 73.22        300
## 91                 34                 62.96        300
## 97                 44                 42.08        300
## 98                 31                 53.42        300
## 105                29                 47.48        300
## 107                36                 33.44        300
## 232                31                 57.38        300
## 234                44                 69.62        300
## 236                16                 73.94        300
## 301               -21                 50.90        300
## 310               -29                 61.70        300
## 318                27                 44.60        300
## 341               -40                 59.90        300
## 343               -16                 64.76        300
## 357               -25                 48.38        300
```

2. Pendekatan Bivariat (2 p/ubah (X dan Y)):

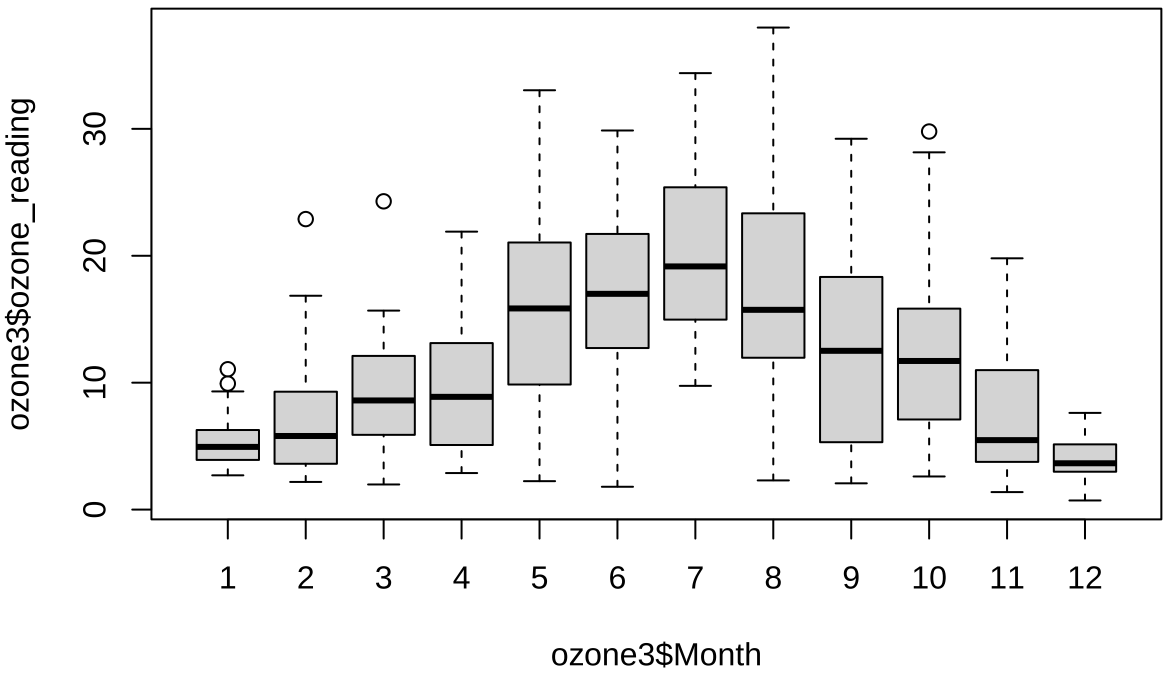
2.1 X ialah kategori dan y berangka

```
par(mfrow=c(1,2))
boxplot(ozone3$ozone_reading~ozone3$Month, main="Plot Kotak Bacaan Ozone Bulanan")
boxplot(ozone3$ozone_reading~ozone3$Day_of_week, main="Plot Kotak Bacaan Ozone Harian")
```



Kesan data pencil dari set bulanan

```
outBiv = boxplot(ozone3$ozone_reading~ozone3$Month)$out
```

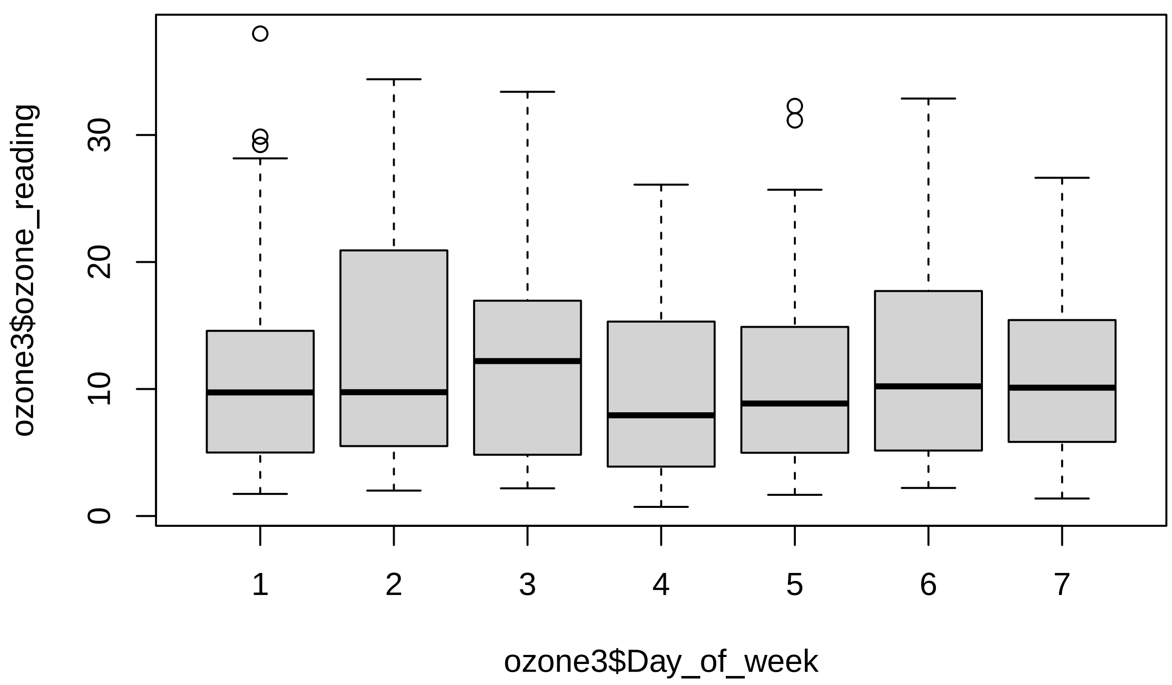


```
out_D = which(ozone3$ozone_reading%in%outBiv)
ozone3[c(out_D),]
```

```
##     Month Day_of_month Day_of_week ozone_reading pressure_height Wind_speed
## 30      1           30           5         11.06            5790          3
## 31      1           31           6          9.93            5800          2
## 58      2           27           5         22.89            5740          3
## 77      3           17           3         24.29            5760          3
## 280    10            6           3         29.79            5890          5
##     Humidity Temperature_Sandburg Temperature_ElMonte Inversion_base_height
## 30        28                   63               57.38                   793
## 31        32                   63               60.98                   531
## 58        47                   53               58.82                   885
## 77        60                   70               58.64                   508
## 280       80                   75               71.06                  1049
##     Pressure_gradient Inversion_temperature Visibility
## 30                -15                 65.84        120
## 31                -38                 75.92         40
## 58                 -4                 67.10         80
## 77                  7                 66.56         70
## 280               -10                 78.98         50
```

Kesan data pencil dari set harian

```
outBiv = boxplot(ozone3$ozone_reading~ozone3$Day_of_week)$out
```



```
out_H = which(ozone3$ozone_reading%in%outBiv)
ozone3[c(out_H),]
```

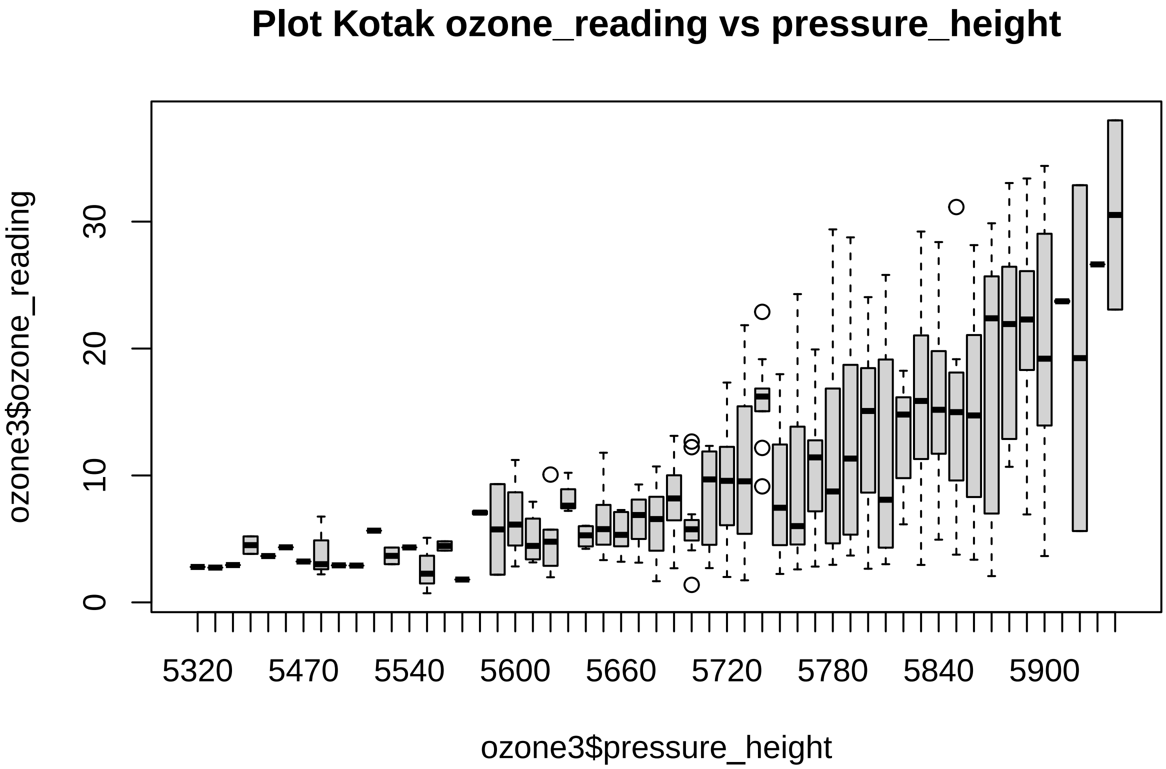
```
##     Month Day_of_month Day_of_week ozone_reading pressure_height Wind_speed
## 135     5           14           5         31.15            5850          4
## 180     6           28           1         29.87            5870          7
## 240     8           27           5         32.28            5900          6
## 243     8           30           1         37.98            5950          5
## 257     9           13           1         29.22            5830          5
##     Humidity Temperature_Sandburg Temperature_ElMonte Inversion_base_height
## 135       76                   78               71.24                  1181
## 180       55                   93               81.68                   646
## 240       71                   87               76.46                   869
## 243       62                   92               82.40                   557
## 257       77                   72               68.72                  1853
##     Pressure_gradient Inversion_temperature Visibility
## 135                50                 79.88         17
## 180                25                 89.24        140
## 240                19                 78.98         17
## 243                 0                 90.68         70
## 257                10                 70.88         70
```

2.2 X ialah berangka dan y berangka

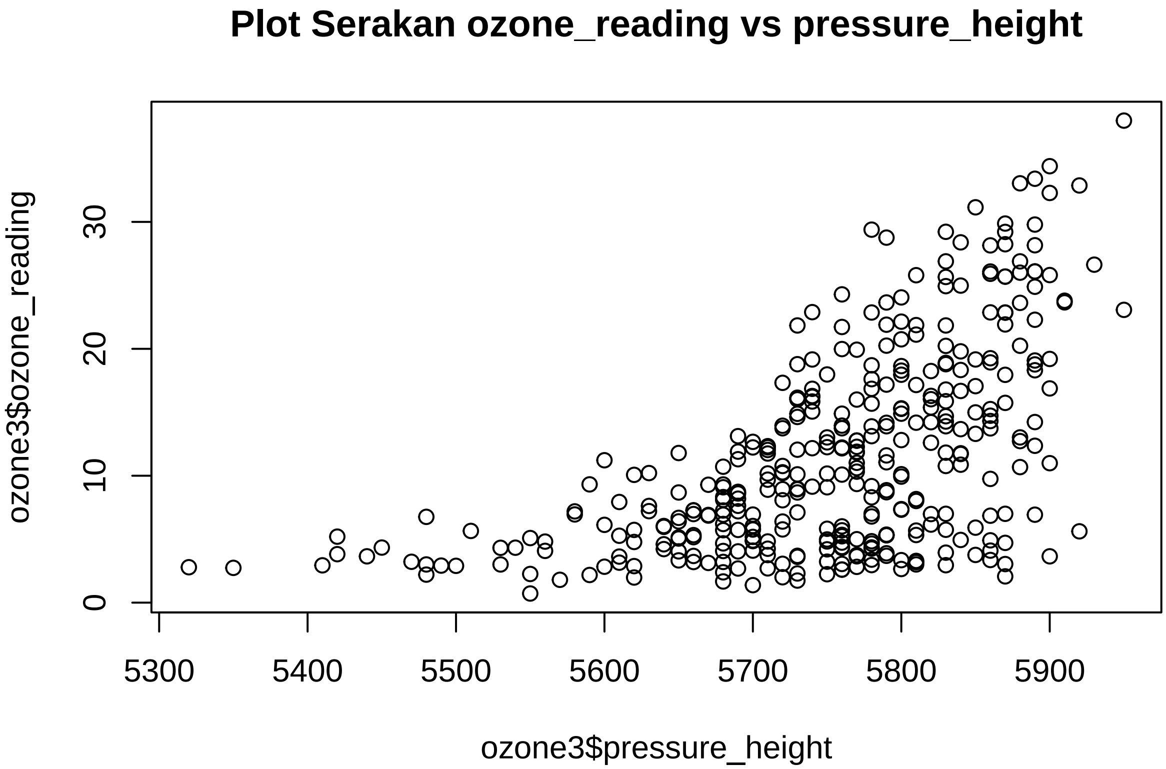
```r
head(ozone3,5)
```

```
##   Month Day_of_month Day_of_week ozone_reading pressure_height Wind_speed
## 1     1            1           4          3.01            5480          8
## 2     1            2           5          3.20            5660          6
## 3     1            3           6          2.70            5710          4
## 4     1            4           7          5.18            5700          3
## 5     1            5           1          5.34            5760          3
##   Humidity Temperature_Sandburg Temperature_ElMonte Inversion_base_height
## 1       20                   30               32.54                  5000
## 2       32                   38               41.36                  1601
## 3       28                   40               38.12                  2693
## 4       37                   45               47.12                   590
## 5       51                   54               45.32                  1450
##   Pressure_gradient Inversion_temperature Visibility
## 1               -15                 30.56        200
## 2               -14                 46.94        300
## 3               -25                 47.66        250
## 4               -24                 55.04        100
## 5                25                 57.02         60
```

```r
boxplot(ozone3$ozone_reading~ozone3$pressure_height, main = "Plot Kotak ozone_reading vs pressure_heigh
```



```
plot(ozone3$pressure_height, ozone3$ozone_reading, main = "Plot Serakan ozone_reading vs pressure_heigh
```



```
x_min = 5400
y_min = 5
```

Kenal pasti data pencil, setkan nilai ambang (threshold) bersesuaian nilai ambang bawah (low threshold)

```
x_max = 5900
y_max = 35
```

nilai ambang atas (high threshold)

```
outlier_MinT = ozone3[ozone3$pressure_height < x_min & ozone3$ozone_reading < y_min,]
outlier_MinT
```

kesan data pencil dari data asal

```
##    Month Day_of_month Day_of_week ozone_reading pressure_height Wind_speed
## 37     2            6           5          2.74            5350          7
## 63     3            3           3          2.79            5320         11
##    Humidity Temperature_Sandburg Temperature_ElMonte Inversion_base_height
## 37       62                   30               32.54                  1341
## 63       45                   25               27.68                  5000
##    Pressure_gradient Inversion_temperature Visibility
## 37                18                 45.86         60
## 63                39                 27.50        200
```

```
outlier_MaxT = ozone3[ozone3$pressure_height > x_max & ozone3$ozone_reading > y_max,]
outlier_MaxT
```

```
##     Month Day_of_month Day_of_week ozone_reading pressure_height Wind_speed
## 243     8           30           1         37.98            5950          5
##     Humidity Temperature_Sandburg Temperature_ElMonte Inversion_base_height
## 243       62                   92                82.4                   557
##     Pressure_gradient Inversion_temperature Visibility
## 243                 0                 90.68         70
```

3. Pendekatan Multivariat

3.1 kes terselia

y = ozone_reading

x = lain pemboleh ubah

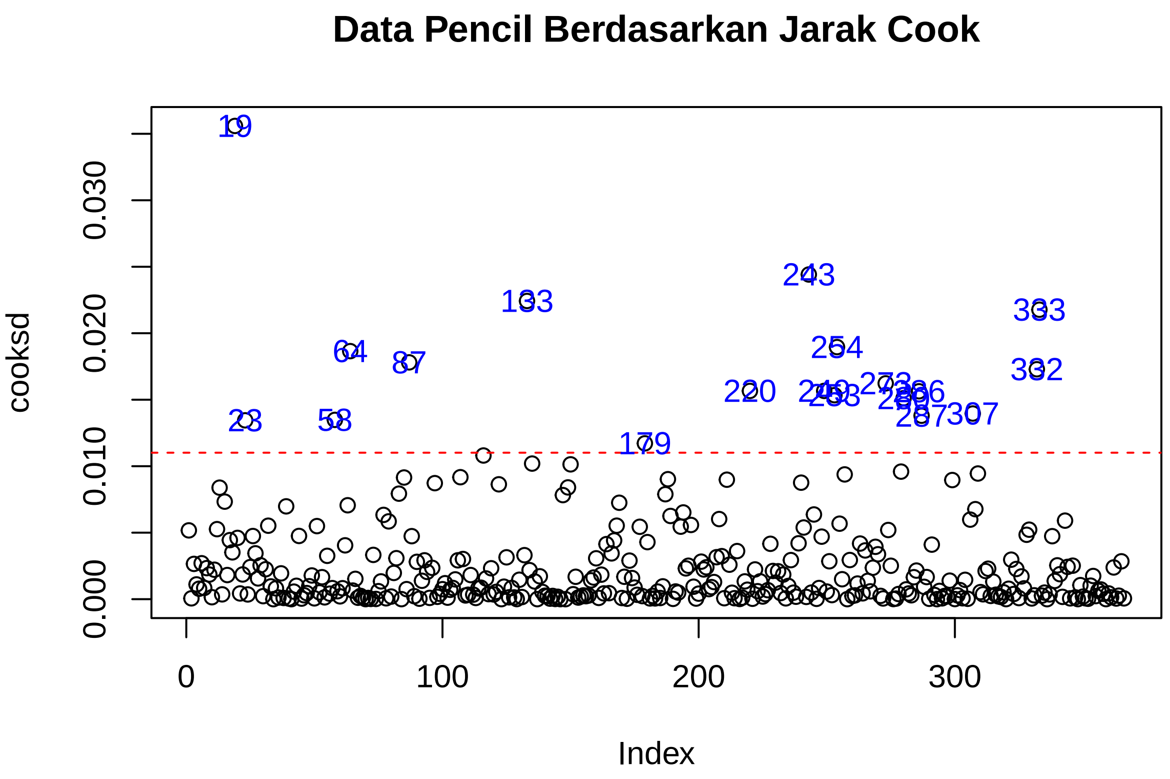
```
model.Reg = lm(ozone_reading~.,data = ozone3)
summary(model.Reg)
```

```
## 
## Call:
## lm(formula = ozone_reading ~ ., data = ozone3)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -12.078  -2.806  -0.095   2.466  13.774 
## 
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)    
## (Intercept)           79.7301501 26.8208664   2.973 0.003155 ** 
## Month                 -0.2912283  0.0772391  -3.770 0.000191 ***
## Day_of_month           0.0117107  0.0260899   0.449 0.653809    
## Day_of_week            0.0082434  0.1139942   0.072 0.942393    
## pressure_height       -0.0168796  0.0050603  -3.336 0.000941 ***
## Wind_speed            -0.1979789  0.1241713  -1.594 0.111741    
## Humidity               0.0592464  0.0175431   3.377 0.000814 ***
## Temperature_Sandburg   0.1595799  0.0516012   3.093 0.002142 ** 
## Temperature_ElMonte    0.5877651  0.0899720   6.533 2.26e-10 ***
## Inversion_base_height -0.0010628  0.0002971  -3.577 0.000396 ***
## Pressure_gradient      0.0118259  0.0106706   1.108 0.268502
## Inversion_temperature -0.1990735  0.0870369  -2.287 0.022773 *
## Visibility            -0.0039045  0.0034562  -1.130 0.259370
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.285 on 353 degrees of freedom
## Multiple R-squared:  0.7161, Adjusted R-squared:  0.7064
## F-statistic: 74.18 on 12 and 353 DF,  p-value: < 2.2e-16
```

```r
cooksd = cooks.distance(model.Reg)
```

Jarak Cook kenal pasti data pencil

```r
plot(cooksd, main="Data Pencil Berdasarkan Jarak Cook")
min_cook = 4*mean(cooksd)
abline(h=(min_cook), col='red', lty=2)

text(x=1:length(cooksd), y=cooksd,
     labels = ifelse(cooksd>min_cook, names(cooksd), ""), col='blue')
```



ekstrak data outlier

```r
outlier_cook = as.numeric(names(cooksd)[cooksd>min_cook])
ozone3[outlier_cook,]
```

```
##     Month Day_of_month Day_of_week ozone_reading pressure_height Wind_speed
## 19      1           19           1          4.07            5680          5
## 23      1           23           5          4.90            5700          5
## 58      2           27           5         22.89            5740          3
## 64      3            4           4          5.20            5420          8
## 87      3           27           6         11.22            5600          6
## 133     5           12           3         33.04            5880          3
## 179     6           27           7         12.73            5880          5
## 220     8            7           6         24.94            5830          4
## 243     8           30           1         37.98            5950          5
## 249     9            5           7         10.12            5800          6
## 253     9            9           4          3.36            5860          5
## 254     9           10           5          2.07            5870          6
## 273     9           29           3          4.60            5640          5
## 280    10            6           3         29.79            5890          5
## 286    10           12           2          7.00            5830          8
## 287    10           13           3         28.15            5860          5
## 307    11            2           2          4.71            5870          6
## 332    11           27           6          3.13            5670          8
## 333    11           28           7          3.05            5760          0
##     Humidity Temperature_Sandburg Temperature_ElMonte Inversion_base_height
## 19        73                   52               56.48                   393
## 23        59                   69               51.08                  3044
## 58        47                   53               58.82                   885
## 64        33                   39               30.20                  5000
## 87        45                   40               41.72                  5000
## 133       80                   80               73.04                   436
## 179       43                   90               73.22                   580
## 220       71                   69               64.04                  5000
## 243       62                   92               82.40                   557
## 249       74                   78               73.22                  2818
## 253       73                   69               66.92                   774
## 254       74                   59               61.88                   134
## 273       93                   63               54.32                  5000
## 280       80                   75               71.06                  1049
## 286       77                   71               67.10                   337
## 287       86                   73               69.80                   492
## 307       58                   68               68.90                  1341
## 332       19                   34               41.00                  5000
## 333       19                   36               38.12                  5000
##     Pressure_gradient Inversion_temperature Visibility
## 19                -68                 69.80         10
## 23                 18                 52.88        150
## 58                 -4                 67.10         80
## 64                 15                 30.02        500
## 87                 38                 46.94        150
## 133                 0                 86.36         40
## 179                 9                 87.26         80
## 220                30                 55.76        100
## 243                 0                 90.68         70
## 249                26                  72.68        70
## 253               -27                  75.56       100
## 254                 0                  77.18        70
## 273                30                  52.70        70
## 280               -10                  78.98        50
## 286               -17                  81.14        20
## 287                -2                  82.22         7
## 307               -42                  73.58       150
## 332               -63                  37.04       150
## 333               -52                  41.00       100
```

3.2 kes tak terselia

```
dataMUS  = read.csv("G:/My Drive/Master-Data-Science/Semester_1/Data_Mining/Data/dataMUS.csv", header=T
dataMUS = dataMUS[-1]
```

hitung jarak Mahalanobis

```
M_dist = mahalanobis(dataMUS, center = colMeans(dataMUS), cov = cov(dataMUS))
```

setkan nilai ambang untuk kesan data pencil, 97.5 persentil untuk taburan khi-kuasa dua.

```
ambang = qchisq(0.975, df=ncol(dataMUS))

outlier_MD = which(M_dist>ambang)

dataMUS[outlier_MD,]
```

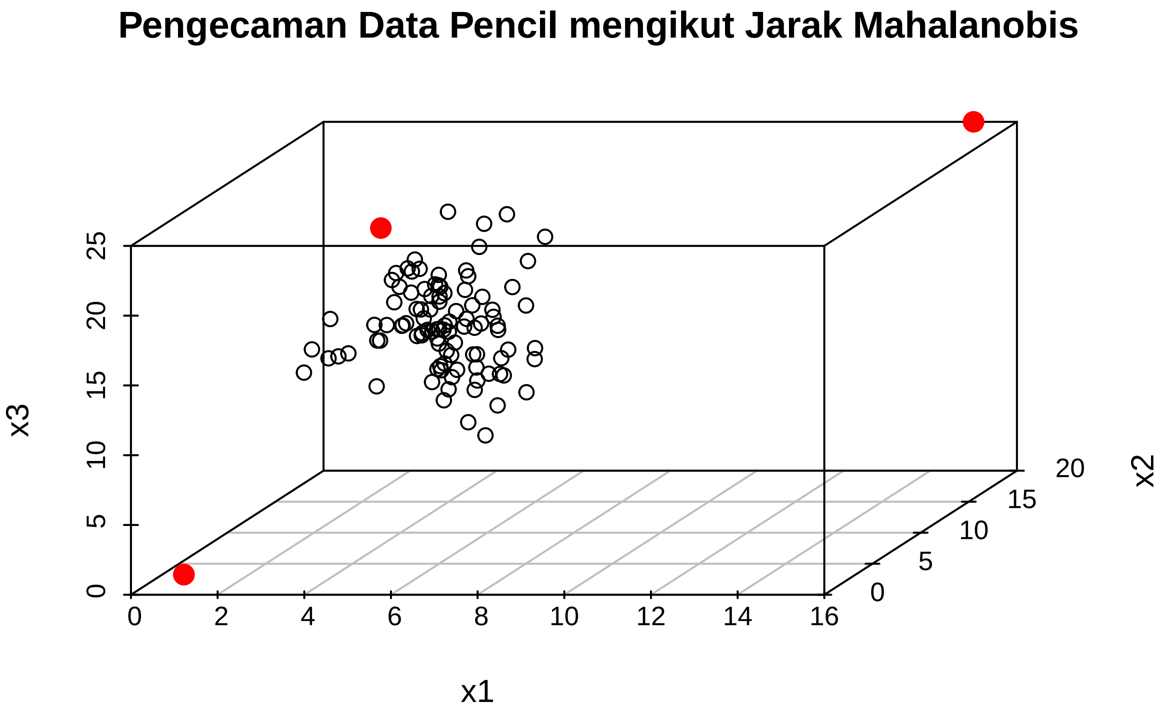
```
##            x1       x2       x3
## 18   2.343545 15.40378 19.43036
## 101 15.000000 20.00000 25.00000
## 102  1.000000  1.00000  1.00000
## 103  1.000000  1.00000  1.00000
```

pengvisualan 3d

```
library(scatterplot3d)
install.packages("scatterplot3d")
```

```
## Warning: package 'scatterplot3d' is in use and will not be installed
```

```
s3d = scatterplot3d(dataMUS, main="Pengecaman Data Pencil mengikut Jarak Mahalanobis")
s3d$points(dataMUS[outlier_MD,], col='red', pch = 16, cex = 1.5)
```


Pengecaman Data Pencil mengikut Jarak Mahalanobis
x3
0
5
10
15
20
25
x1
0
2
4
6
8
10
12
14
16
x2
0
5
10
15
20