Chapter 1 : Data Preprocessing

문제0. Tidyverse, caret, data.table 패키지를 불러오세요.

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
## ✓ ggplot2 3.4.0      ✓ purrr   1.0.0
## ✓ tibble  3.1.8      ✓ dplyr   1.0.10
## ✓ tidyr   1.2.1      ✓ stringr 1.5.0
## ✓ readr   2.1.3      ✓ forcats 0.5.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
library(caret)
```

```
## 필요한 패키지를 로딩중입니다: lattice
##
## 다음의 패키지를 부착합니다: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(data.table)
```

```
##
## 다음의 패키지를 부착합니다: 'data.table'
##
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
##
## The following object is masked from 'package:purrr':
##
##     transpose
```

```
library(magrittr)
```

```
##
## 다음의 패키지를 부착합니다: 'magrittr'
##
## The following object is masked from 'package:purrr':
##
##     set_names
##
## The following object is masked from 'package:tidyr':
##
##     extract
```

문제1. Train, test 데이터를 불러온 뒤 데이터의 구조를 파악하세요.

```
train<-fread('train.csv')
test<-fread('test.csv')
```

```
summary(train)
```

```
##   encounter_id      patient_id      hospital_id         age
##  Min.   :    11   Min.   :     4   Min.   :  2.0   Min.   :16.00
##  1st Qu.: 33146   1st Qu.: 32910   1st Qu.: 47.0   1st Qu.:53.00
##  Median : 65870   Median : 65333   Median :109.0   Median :65.00
##  Mean   : 65660   Mean   : 65451   Mean   :105.3   Mean   :62.41
##  3rd Qu.: 98744   3rd Qu.: 98037   3rd Qu.:161.0   3rd Qu.:75.00
##  Max.   :131049   Max.   :131048   Max.   :204.0   Max.   :89.00
##                                                    NA's   :392
##       bmi        elective_surgery  ethnicity           gender
##  Min.   :14.84   Min.   :0.0000   Length:8400        Length:8400
##  1st Qu.:23.62   1st Qu.:0.0000   Class :character   Class :character
##  Median :27.71   Median :0.0000   Mode  :character   Mode  :character
##  Mean   :29.22   Mean   :0.1807
##  3rd Qu.:33.11   3rd Qu.:0.0000
##  Max.   :67.81   Max.   :1.0000
##  NA's   :330
##      height      icu_admit_source       icu_id      icu_stay_type
##  Min.   :137.2   Length:8400        Min.   : 82.0   Length:8400
##  1st Qu.:162.6   Class :character   1st Qu.:362.0   Class :character
##  Median :170.1   Mode  :character   Median :498.0   Mode  :character
##  Mean   :169.7                      Mean   :503.7
##  3rd Qu.:177.8                      3rd Qu.:678.0
##  Max.   :195.6                      Max.   :927.0
##  NA's   :120
##    icu_type         pre_icu_los_days       weight       apache_2_diagnosis
##  Length:8400        Min.   :-5.22778   Min.   : 38.60   Min.   :101.0
##  Class :character   1st Qu.: 0.03403   1st Qu.: 67.00   1st Qu.:113.0
##  Mode  :character   Median : 0.14097   Median : 80.50   Median :122.0
##                     Mean   : 0.87216   Mean   : 84.17   Mean   :183.9
##                     3rd Qu.: 0.41181   3rd Qu.: 97.50   3rd Qu.:301.0
##                     Max.   :64.94861   Max.   :186.00   Max.   :308.0
##                                        NA's   :265      NA's   :159
##  apache_3j_diagnosis apache_post_operative   arf_apache      gcs_eyes_apache
##  Min.   :   0.02     Min.   :0.0000        Min.   :0.00000   Min.   :1.000
##  1st Qu.: 203.01     1st Qu.:0.0000        1st Qu.:0.00000   1st Qu.:3.000
##  Median : 409.02     Median :0.0000        Median :0.00000   Median :4.000
##  Mean   : 552.36     Mean   :0.1971        Mean   :0.02591   Mean   :3.459
##  3rd Qu.: 703.03     3rd Qu.:0.0000        3rd Qu.:0.00000   3rd Qu.:4.000
##  Max.   :2201.05     Max.   :1.0000        Max.   :1.00000   Max.   :4.000
##  NA's   :94                                NA's   :63        NA's   :169
##  gcs_motor_apache gcs_unable_apache gcs_verbal_apache heart_rate_apache
##  Min.   :1.000    Min.   :0.0000    Min.   :1.000     Min.   : 30.0
##  1st Qu.:6.000    1st Qu.:0.0000    1st Qu.:4.000     1st Qu.: 87.0
##  Median :6.000    Median :0.0000    Median :5.000     Median :104.0
##  Mean   :5.475    Mean   :0.0101    Mean   :3.991     Mean   :100.1
##  3rd Qu.:6.000    3rd Qu.:0.0000    3rd Qu.:5.000     3rd Qu.:120.0
##  Max.   :6.000    Max.   :1.0000    Max.   :5.000     Max.   :178.0
##  NA's   :169      NA's   :85        NA's   :169       NA's   :74
##  intubated_apache   map_apache     resprate_apache  temp_apache
##  Min.   :0.0000   Min.   : 40.00   Min.   : 4.00   Min.   :32.10
##  1st Qu.:0.0000   1st Qu.: 54.00   1st Qu.:11.00   1st Qu.:36.20
##  Median :0.0000   Median : 66.00   Median :28.00   Median :36.50
##  Mean   :0.1541   Mean   : 87.97   Mean   :26.09   Mean   :36.42
##  3rd Qu.:0.0000   3rd Qu.:125.00   3rd Qu.:36.00   3rd Qu.:36.70
##  Max.   :1.0000   Max.   :200.00   Max.   :60.00   Max.   :39.70
```

```
##  NA's   :63       NA's   :83       NA's   :111     NA's   :388
##  ventilated_apache d1_diasbp_max    d1_diasbp_min   d1_diasbp_noninvasive_max
##  Min.   :0.0000    Min.   : 46.00   Min.   :13.00   Min.   : 46.00
##  1st Qu.:0.0000    1st Qu.: 75.00   1st Qu.:42.00   1st Qu.: 75.00
##  Median :0.0000    Median : 86.00   Median :50.00   Median : 87.00
##  Mean   :0.3306    Mean   : 88.51   Mean   :50.19   Mean   : 88.67
##  3rd Qu.:1.0000    3rd Qu.: 99.00   3rd Qu.:58.00   3rd Qu.: 99.00
##  Max.   :1.0000    Max.   :165.00   Max.   :90.00   Max.   :165.00
##  NA's   :63        NA's   :13       NA's   :13      NA's   :98
##  d1_diasbp_noninvasive_min d1_heartrate_max d1_heartrate_min   d1_mbp_max
##  Min.   :13.00             Min.   : 58.0    Min.   :  0.00   Min.   : 60.0
##  1st Qu.:42.00             1st Qu.: 88.0    1st Qu.: 60.00   1st Qu.: 90.0
##  Median :50.00             Median :101.0    Median : 69.00   Median :102.0
##  Mean   :50.28             Mean   :103.2    Mean   : 70.21   Mean   :104.7
##  3rd Qu.:58.00             3rd Qu.:117.0    3rd Qu.: 80.00   3rd Qu.:116.0
##  Max.   :90.00             Max.   :177.0    Max.   :155.00   Max.   :184.0
##  NA's   :98                NA's   :10       NA's   :10       NA's   :16
##    d1_mbp_min    d1_mbp_noninvasive_max d1_mbp_noninvasive_min d1_resprate_max
##  Min.   : 22.0   Min.   : 60.0          Min.   : 22.00         Min.   :14.00
##  1st Qu.: 55.0   1st Qu.: 90.0          1st Qu.: 55.00         1st Qu.:22.00
##  Median : 64.0   Median :102.0          Median : 64.00         Median :27.00
##  Mean   : 64.8   Mean   :104.6          Mean   : 64.92         Mean   :28.98
##  3rd Qu.: 74.0   3rd Qu.:116.0          3rd Qu.: 74.00         3rd Qu.:32.00
##  Max.   :112.0   Max.   :181.0          Max.   :112.00         Max.   :92.00
##  NA's   :16      NA's   :143            NA's   :143            NA's   :33
##  d1_resprate_min  d1_spo2_max      d1_spo2_min      d1_sysbp_max
##  Min.   : 0.0    Min.   : 31.00   Min.   :  0.00   Min.   : 90.0
##  1st Qu.:10.0    1st Qu.: 99.00   1st Qu.: 89.00   1st Qu.:130.0
##  Median :13.0    Median :100.00   Median : 92.00   Median :146.0
##  Mean   :12.9    Mean   : 99.26   Mean   : 90.34   Mean   :148.1
##  3rd Qu.:16.0    3rd Qu.:100.00   3rd Qu.: 95.00   3rd Qu.:164.0
##  Max.   :53.0    Max.   :100.00   Max.   :100.00   Max.   :232.0
##  NA's   :33      NA's   :32       NA's   :32       NA's   :12
##   d1_sysbp_min    d1_sysbp_noninvasive_max d1_sysbp_noninvasive_min
##  Min.   : 41.00   Min.   : 90              Min.   : 41.03
##  1st Qu.: 83.00   1st Qu.:130              1st Qu.: 83.00
##  Median : 95.00   Median :146              Median : 95.00
##  Mean   : 96.72   Mean   :148              Mean   : 96.84
##  3rd Qu.:109.00   3rd Qu.:164              3rd Qu.:109.00
##  Max.   :160.00   Max.   :232              Max.   :160.00
##  NA's   :12       NA's   :96               NA's   :96
##   d1_temp_max    d1_temp_min    h1_diasbp_max   h1_diasbp_min
##  Min.   :35.1   Min.   :31.89   Min.   : 37.0   Min.   : 22.00
##  1st Qu.:36.9   1st Qu.:36.10   1st Qu.: 62.0   1st Qu.: 51.00
##  Median :37.2   Median :36.40   Median : 73.0   Median : 61.00
##  Mean   :37.3   Mean   :36.27   Mean   : 75.1   Mean   : 62.64
##  3rd Qu.:37.6   3rd Qu.:36.66   3rd Qu.: 86.0   3rd Qu.: 73.00
##  Max.   :39.9   Max.   :37.80   Max.   :143.0   Max.   :113.00
##  NA's   :226    NA's   :226     NA's   :317     NA's   :317
##  h1_diasbp_noninvasive_max h1_diasbp_noninvasive_min h1_heartrate_max
##  Min.   : 37.00            Min.   : 22.00            Min.   : 46.00
##  1st Qu.: 63.00            1st Qu.: 52.00            1st Qu.: 77.00
##  Median : 73.00            Median : 62.00            Median : 90.00
##  Mean   : 75.57            Mean   : 63.06            Mean   : 92.64
##  3rd Qu.: 86.00            3rd Qu.: 73.00            3rd Qu.:106.00
##  Max.   :144.00            Max.   :114.00            Max.   :164.00
```

```
##  NA's   :645              NA's   :645              NA's   :245
##  h1_heartrate_min   h1_mbp_max       h1_mbp_min     h1_mbp_noninvasive_max
##  Min.   : 36.00   Min.   : 49.00   Min.   : 32.00   Min.   : 49.00
##  1st Qu.: 69.00   1st Qu.: 77.00   1st Qu.: 66.00   1st Qu.: 77.00
##  Median : 82.00   Median : 89.00   Median : 78.00   Median : 89.00
##  Mean   : 83.96   Mean   : 91.34   Mean   : 79.15   Mean   : 91.32
##  3rd Qu.: 97.00   3rd Qu.:103.00   3rd Qu.: 92.00   3rd Qu.:103.00
##  Max.   :144.00   Max.   :165.00   Max.   :138.00   Max.   :163.00
##  NA's   :245      NA's   :401      NA's   :401      NA's   :813
##  h1_mbp_noninvasive_min h1_resprate_max h1_resprate_min   h1_spo2_max
##  Min.   : 32.00         Min.   :10.0    Min.   :  0.00   Min.   :  2.00
##  1st Qu.: 66.00         1st Qu.:18.0    1st Qu.: 14.00   1st Qu.: 97.00
##  Median : 78.00         Median :21.0    Median : 16.00   Median : 99.00
##  Mean   : 79.54         Mean   :22.7    Mean   : 17.27   Mean   : 98.04
##  3rd Qu.: 92.00         3rd Qu.:26.0    3rd Qu.: 20.00   3rd Qu.:100.00
##  Max.   :138.00         Max.   :59.0    Max.   :100.00   Max.   :100.00
##  NA's   :813            NA's   :400     NA's   :400      NA's   :381
##   h1_spo2_min      h1_sysbp_max    h1_sysbp_min h1_sysbp_noninvasive_max
##  Min.   :  0.00   Min.   : 75.0   Min.   : 53   Min.   : 75.0
##  1st Qu.: 94.00   1st Qu.:113.0   1st Qu.: 97   1st Qu.:113.0
##  Median : 96.00   Median :130.0   Median :114   Median :130.0
##  Mean   : 95.11   Mean   :132.8   Mean   :116   Mean   :132.6
##  3rd Qu.: 99.00   3rd Qu.:150.0   3rd Qu.:133   3rd Qu.:150.0
##  Max.   :100.00   Max.   :223.0   Max.   :194   Max.   :223.0
##  NA's   :381      NA's   :316     NA's   :316   NA's   :645
##  h1_sysbp_noninvasive_min d1_glucose_max  d1_glucose_min  d1_potassium_max
##  Min.   : 53.0            Min.   : 73.0   Min.   : 33.0   Min.   :2.800
##  1st Qu.: 97.0            1st Qu.:117.0   1st Qu.: 91.0   1st Qu.:3.800
##  Median :114.0            Median :151.0   Median :108.0   Median :4.200
##  Mean   :116.1            Mean   :175.5   Mean   :114.9   Mean   :4.236
##  3rd Qu.:134.0            3rd Qu.:203.0   3rd Qu.:131.0   3rd Qu.:4.600
##  Max.   :195.0            Max.   :611.0   Max.   :288.0   Max.   :7.000
##  NA's   :645              NA's   :493     NA's   :493     NA's   :839
##  d1_potassium_min apache_4a_hospital_death_prob apache_4a_icu_death_prob
##  Min.   :2.400    Min.   :-1.000                Min.   :-1.0000
##  1st Qu.:3.600    1st Qu.: 0.020                1st Qu.: 0.0100
##  Median :3.900    Median : 0.050                Median : 0.0200
##  Mean   :3.928    Mean   : 0.086                Mean   : 0.0424
##  3rd Qu.:4.300    3rd Qu.: 0.130                3rd Qu.: 0.0700
##  Max.   :5.800    Max.   : 0.960                Max.   : 0.9500
##  NA's   :839      NA's   :736                   NA's   :736
##       aids           cirrhosis       diabetes_mellitus hepatic_failure
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.000     Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.000     1st Qu.:0.00000
##  Median :0.00000   Median :0.00000   Median :0.000     Median :0.00000
##  Mean   :0.00072   Mean   :0.01343   Mean   :0.225     Mean   :0.01187
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.000     3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.000     Max.   :1.00000
##  NA's   :63        NA's   :63        NA's   :63        NA's   :63
##  immunosuppression    leukemia          lymphoma
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.00000   Median :0.00000   Median :0.00000
##  Mean   :0.02903   Mean   :0.00804   Mean   :0.00396
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
```

```
##  NA's   :63        NA's   :63        NA's   :63
##  solid_tumor_with_metastasis apache_3j_bodysystem apache_2_bodysystem
##  Min.   :0.00000             Length:8400          Length:8400
##  1st Qu.:0.00000             Class :character     Class :character
##  Median :0.00000             Mode  :character     Mode  :character
##  Mean   :0.02027
##  3rd Qu.:0.00000
##  Max.   :1.00000
##  NA's   :63
##    V84          hospital_death
##  Mode:logical   Min.   :0.00000
##  NA's:8400      1st Qu.:0.00000
##                 Median :0.00000
##                 Mean   :0.08429
##                 3rd Qu.:0.00000
##                 Max.   :1.00000
##
```

```
summary(test)
```

```
##   encounter_id      patient_id      hospital_id         age
##  Min.   :    15   Min.   :     1   Min.   :  2.0   Min.   :16.00
##  1st Qu.: 31761   1st Qu.: 31866   1st Qu.: 49.0   1st Qu.:52.00
##  Median : 66152   Median : 64123   Median :111.0   Median :65.00
##  Mean   : 65868   Mean   : 65263   Mean   :105.3   Mean   :62.31
##  3rd Qu.: 99774   3rd Qu.: 98511   3rd Qu.:161.0   3rd Qu.:75.00
##  Max.   :131032   Max.   :130988   Max.   :204.0   Max.   :89.00
##                                                    NA's   :159
##       bmi        elective_surgery  ethnicity           gender
##  Min.   :14.84   Min.   :0.0000   Length:3600        Length:3600
##  1st Qu.:23.60   1st Qu.:0.0000   Class :character   Class :character
##  Median :27.77   Median :0.0000   Mode  :character   Mode  :character
##  Mean   :29.19   Mean   :0.1914
##  3rd Qu.:33.20   3rd Qu.:0.0000
##  Max.   :67.81   Max.   :1.0000
##  NA's   :124
##      height      icu_admit_source       icu_id      icu_stay_type
##  Min.   :137.2   Length:3600        Min.   : 82.0   Length:3600
##  1st Qu.:162.6   Class :character   1st Qu.:369.0   Class :character
##  Median :170.0   Mode  :character   Median :498.0   Mode  :character
##  Mean   :169.5                      Mean   :508.2
##  3rd Qu.:177.8                      3rd Qu.:683.0
##  Max.   :195.6                      Max.   :927.0
##  NA's   :39
##    icu_type         pre_icu_los_days       weight       apache_2_diagnosis
##  Length:3600        Min.   :-1.77569   Min.   : 38.60   Min.   :101.0
##  Class :character   1st Qu.: 0.03681   1st Qu.: 66.00   1st Qu.:113.0
##  Mode  :character   Median : 0.14306   Median : 81.00   Median :122.0
##                     Mean   : 0.90820   Mean   : 83.99   Mean   :185.1
##                     3rd Qu.: 0.44028   3rd Qu.: 97.30   3rd Qu.:301.0
##                     Max.   :57.64931   Max.   :186.00   Max.   :308.0
##                                        NA's   :103      NA's   :72
##  apache_3j_diagnosis apache_post_operative   arf_apache      gcs_eyes_apache
##  Min.   :   0.04     Min.   :0.0000        Min.   :0.00000   Min.   :1.000
##  1st Qu.: 203.76     1st Qu.:0.0000        1st Qu.:0.00000   1st Qu.:3.000
##  Median : 410.01     Median :0.0000        Median :0.00000   Median :4.000
##  Mean   : 564.87     Mean   :0.2083        Mean   :0.02687   Mean   :3.477
##  3rd Qu.: 704.01     3rd Qu.:0.0000        3rd Qu.:0.00000   3rd Qu.:4.000
##  Max.   :2201.05     Max.   :1.0000        Max.   :1.00000   Max.   :4.000
##  NA's   :48                                NA's   :27        NA's   :77
##  gcs_motor_apache gcs_unable_apache gcs_verbal_apache heart_rate_apache
##  Min.   :1.000    Min.   :0.00000   Min.   :1.000     Min.   : 30.0
##  1st Qu.:6.000    1st Qu.:0.00000   1st Qu.:4.000     1st Qu.: 88.0
##  Median :6.000    Median :0.00000   Median :5.000     Median :104.0
##  Mean   :5.462    Mean   :0.00956   Mean   :4.006     Mean   :100.2
##  3rd Qu.:6.000    3rd Qu.:0.00000   3rd Qu.:5.000     3rd Qu.:120.0
##  Max.   :6.000    Max.   :1.00000   Max.   :5.000     Max.   :178.0
##  NA's   :77       NA's   :43        NA's   :77        NA's   :37
##  intubated_apache   map_apache     resprate_apache  temp_apache
##  Min.   :0.0000   Min.   : 40.00   Min.   : 4.00   Min.   :32.10
##  1st Qu.:0.0000   1st Qu.: 53.00   1st Qu.:11.00   1st Qu.:36.16
##  Median :0.0000   Median : 67.00   Median :28.00   Median :36.50
##  Mean   :0.1551   Mean   : 87.84   Mean   :25.95   Mean   :36.40
##  3rd Qu.:0.0000   3rd Qu.:124.00   3rd Qu.:37.00   3rd Qu.:36.70
##  Max.   :1.0000   Max.   :200.00   Max.   :60.00   Max.   :39.70
```

```
##  NA's   :27       NA's   :39       NA's   :51      NA's   :151
##  ventilated_apache d1_diasbp_max    d1_diasbp_min   d1_diasbp_noninvasive_max
##  Min.   :0.0000    Min.   : 46.00   Min.   :13.00   Min.   : 46.00
##  1st Qu.:0.0000    1st Qu.: 76.00   1st Qu.:42.00   1st Qu.: 76.00
##  Median :0.0000    Median : 87.00   Median :50.00   Median : 87.00
##  Mean   :0.3261    Mean   : 88.65   Mean   :50.05   Mean   : 88.79
##  3rd Qu.:1.0000    3rd Qu.: 99.00   3rd Qu.:58.00   3rd Qu.: 99.00
##  Max.   :1.0000    Max.   :165.00   Max.   :90.00   Max.   :165.00
##  NA's   :27        NA's   :8        NA's   :8       NA's   :43
##  d1_diasbp_noninvasive_min d1_heartrate_max d1_heartrate_min   d1_mbp_max
##  Min.   :13.00             Min.   : 58.0    Min.   :  0.0    Min.   : 60.0
##  1st Qu.:42.00             1st Qu.: 87.0    1st Qu.: 60.0    1st Qu.: 91.0
##  Median :50.00             Median :101.0    Median : 70.0    Median :102.0
##  Mean   :50.12             Mean   :102.9    Mean   : 70.1    Mean   :104.9
##  3rd Qu.:58.00             3rd Qu.:116.0    3rd Qu.: 80.0    3rd Qu.:116.0
##  Max.   :90.00             Max.   :177.0    Max.   :142.0    Max.   :184.0
##  NA's   :43                NA's   :9        NA's   :9        NA's   :10
##    d1_mbp_min     d1_mbp_noninvasive_max d1_mbp_noninvasive_min d1_resprate_max
##  Min.   : 22.00   Min.   : 60.0          Min.   : 22.00         Min.   :14.00
##  1st Qu.: 54.00   1st Qu.: 91.0          1st Qu.: 54.00         1st Qu.:22.00
##  Median : 64.00   Median :102.0          Median : 64.00         Median :27.00
##  Mean   : 64.62   Mean   :104.7          Mean   : 64.69         Mean   :29.13
##  3rd Qu.: 75.00   3rd Qu.:116.0          3rd Qu.: 75.00         3rd Qu.:33.00
##  Max.   :112.00   Max.   :181.0          Max.   :112.00         Max.   :92.00
##  NA's   :10       NA's   :56             NA's   :56             NA's   :13
##  d1_resprate_min   d1_spo2_max      d1_spo2_min      d1_sysbp_max
##  Min.   :  0.00   Min.   : 67.00   Min.   :  0.00   Min.   : 90.0
##  1st Qu.: 10.00   1st Qu.: 99.00   1st Qu.: 89.00   1st Qu.:130.0
##  Median : 13.00   Median :100.00   Median : 92.00   Median :146.0
##  Mean   : 12.84   Mean   : 99.23   Mean   : 90.11   Mean   :148.8
##  3rd Qu.: 16.00   3rd Qu.:100.00   3rd Qu.: 95.00   3rd Qu.:164.0
##  Max.   :100.00   Max.   :100.00   Max.   :100.00   Max.   :232.0
##  NA's   :13       NA's   :12       NA's   :12       NA's   :7
##   d1_sysbp_min    d1_sysbp_noninvasive_max d1_sysbp_noninvasive_min
##  Min.   : 41.00   Min.   : 90.0            Min.   : 41.03
##  1st Qu.: 83.00   1st Qu.:130.0            1st Qu.: 83.00
##  Median : 95.00   Median :146.0            Median : 96.00
##  Mean   : 96.85   Mean   :148.6            Mean   : 96.88
##  3rd Qu.:110.00   3rd Qu.:164.0            3rd Qu.:110.00
##  Max.   :160.00   Max.   :232.0            Max.   :160.00
##  NA's   :7        NA's   :41               NA's   :41
##   d1_temp_max     d1_temp_min    h1_diasbp_max    h1_diasbp_min
##  Min.   :35.10   Min.   :31.89   Min.   : 37.00   Min.   : 22.00
##  1st Qu.:36.80   1st Qu.:36.10   1st Qu.: 62.00   1st Qu.: 51.00
##  Median :37.11   Median :36.40   Median : 73.00   Median : 61.00
##  Mean   :37.28   Mean   :36.26   Mean   : 75.47   Mean   : 62.72
##  3rd Qu.:37.60   3rd Qu.:36.61   3rd Qu.: 87.00   3rd Qu.: 73.00
##  Max.   :39.90   Max.   :37.80   Max.   :143.00   Max.   :113.00
##  NA's   :78      NA's   :78      NA's   :142      NA's   :142
##  h1_diasbp_noninvasive_max h1_diasbp_noninvasive_min h1_heartrate_max
##  Min.   : 37.00            Min.   : 22.00            Min.   : 46.00
##  1st Qu.: 62.00            1st Qu.: 52.00            1st Qu.: 77.00
##  Median : 74.00            Median : 62.00            Median : 90.00
##  Mean   : 75.78            Mean   : 63.13            Mean   : 92.32
##  3rd Qu.: 87.00            3rd Qu.: 73.00            3rd Qu.:106.00
##  Max.   :144.00            Max.   :114.00            Max.   :164.00
```

```
##  NA's   :285              NA's   :285              NA's   :106
##  h1_heartrate_min   h1_mbp_max       h1_mbp_min     h1_mbp_noninvasive_max
##  Min.   : 36.00   Min.   : 49.00   Min.   : 32.00   Min.   : 49.0
##  1st Qu.: 69.00   1st Qu.: 76.00   1st Qu.: 66.00   1st Qu.: 76.0
##  Median : 82.00   Median : 89.00   Median : 78.00   Median : 89.0
##  Mean   : 83.54   Mean   : 91.54   Mean   : 79.08   Mean   : 91.5
##  3rd Qu.: 97.00   3rd Qu.:104.00   3rd Qu.: 91.00   3rd Qu.:104.0
##  Max.   :144.00   Max.   :165.00   Max.   :138.00   Max.   :163.0
##  NA's   :106      NA's   :169      NA's   :169      NA's   :336
##  h1_mbp_noninvasive_min h1_resprate_max h1_resprate_min   h1_spo2_max
##  Min.   : 32.00         Min.   :10.00   Min.   :  0.00   Min.   :  0
##  1st Qu.: 66.00         1st Qu.:18.00   1st Qu.: 13.00   1st Qu.: 97
##  Median : 78.00         Median :21.00   Median : 16.00   Median : 99
##  Mean   : 79.39         Mean   :22.95   Mean   : 17.26   Mean   : 98
##  3rd Qu.: 92.00         3rd Qu.:26.00   3rd Qu.: 20.00   3rd Qu.:100
##  Max.   :138.00         Max.   :59.00   Max.   :125.00   Max.   :100
##  NA's   :336            NA's   :170     NA's   :170      NA's   :166
##   h1_spo2_min      h1_sysbp_max    h1_sysbp_min   h1_sysbp_noninvasive_max
##  Min.   :  0.00   Min.   : 75.0   Min.   : 53.0   Min.   : 75.0
##  1st Qu.: 94.00   1st Qu.:113.0   1st Qu.: 97.0   1st Qu.:113.0
##  Median : 96.00   Median :131.0   Median :114.0   Median :130.0
##  Mean   : 95.01   Mean   :133.4   Mean   :116.3   Mean   :133.2
##  3rd Qu.: 98.75   3rd Qu.:151.0   3rd Qu.:134.0   3rd Qu.:151.0
##  Max.   :100.00   Max.   :223.0   Max.   :194.0   Max.   :223.0
##  NA's   :166      NA's   :141     NA's   :141     NA's   :284
##  h1_sysbp_noninvasive_min d1_glucose_max  d1_glucose_min  d1_potassium_max
##  Min.   : 53.0            Min.   : 73.0   Min.   : 33.0   Min.   :2.800
##  1st Qu.: 98.0            1st Qu.:118.0   1st Qu.: 90.0   1st Qu.:3.800
##  Median :114.0            Median :150.0   Median :107.0   Median :4.200
##  Mean   :116.4            Mean   :174.9   Mean   :113.3   Mean   :4.246
##  3rd Qu.:134.0            3rd Qu.:202.0   3rd Qu.:130.0   3rd Qu.:4.600
##  Max.   :195.0            Max.   :611.0   Max.   :288.0   Max.   :7.000
##  NA's   :284              NA's   :220     NA's   :220     NA's   :354
##  d1_potassium_min apache_4a_hospital_death_prob apache_4a_icu_death_prob
##  Min.   :2.400    Min.   :-1.00000              Min.   :-1.00000
##  1st Qu.:3.600    1st Qu.: 0.02000              1st Qu.: 0.01000
##  Median :3.900    Median : 0.05000              Median : 0.02000
##  Mean   :3.926    Mean   : 0.08433              Mean   : 0.04219
##  3rd Qu.:4.300    3rd Qu.: 0.13000              3rd Qu.: 0.06000
##  Max.   :5.800    Max.   : 0.96000              Max.   : 0.93000
##  NA's   :354      NA's   :290                   NA's   :290
##       aids           cirrhosis       diabetes_mellitus hepatic_failure
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.0000    Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.0000    1st Qu.:0.00000
##  Median :0.00000   Median :0.00000   Median :0.0000    Median :0.00000
##  Mean   :0.00056   Mean   :0.01903   Mean   :0.2317    Mean   :0.01231
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.0000    3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.0000    Max.   :1.00000
##  NA's   :27        NA's   :27        NA's   :27        NA's   :27
##  immunosuppression    leukemia          lymphoma
##  Min.   :0.00000   Min.   :0.000000   Min.   :0.000000
##  1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.000000
##  Median :0.00000   Median :0.000000   Median :0.000000
##  Mean   :0.02967   Mean   :0.005598   Mean   :0.003918
##  3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:0.000000
##  Max.   :1.00000   Max.   :1.000000   Max.   :1.000000
```

```
##  NA's   :27        NA's   :27         NA's   :27
##  solid_tumor_with_metastasis apache_3j_bodysystem apache_2_bodysystem
##  Min.   :0.00000             Length:3600          Length:3600
##  1st Qu.:0.00000             Class :character     Class :character
##  Median :0.00000             Mode  :character     Mode  :character
##  Mean   :0.02239
##  3rd Qu.:0.00000
##  Max.   :1.00000
##  NA's   :27
##    V84          hospital_death
##  Mode:logical   Min.   :0.00000
##  NA's:3600      1st Qu.:0.00000
##                 Median :0.00000
##                 Mean   :0.08028
##                 3rd Qu.:0.00000
##                 Max.   :1.00000
##
```

```
#str(train)
#str(test)
```

```
library(VIM)
```

```
## 필요한 패키지를 로딩중입니다: colorspace
```

```
## 필요한 패키지를 로딩중입니다: grid
```

```
## VIM is ready to use.
```

```
## Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues
```

```
##
## 다음의 패키지를 부착합니다: 'VIM'
```

```
## The following object is masked from 'package:datasets':
##
##     sleep
```

```
train_na<-train %>% select_if(colSums(is.na(.))>0)
test_na<-test %>% select_if(colSums(is.na(.))>0)

#aggr(train_na,prop=F,numbers=T,col=c('lavender','lavenderblush'),cex.axis=0.77)

#aggr(test_na,prop=F,numbers=T,col=c('lavender','lavenderblush'),cex.axis=0.77)
```

문제2. *id, icu*, noninvasive_가 변수명에 포함된 Column은 불필요하므로 제거해주세요.

(HINT2) dplyr 패키지 내 contains() 함수를 이용하면 좋습니다.

https://zorba78.github.io/cnu-r-programming-lecture-note/dplyr.html (https://zorba78.github.io/cnu-r-programming-lecture-note/dplyr.html)

```
train %<>%
  select(-contains(c('_id','icu_','noninvasive_')))
```

문제3. V84 Column을 제거하고, bmi, height, weight 변수에 대해서 소수점 첫째자리에서 반올림하세요.

```
train %<>%
  select(-V84) %>%
  mutate_at(vars(bmi, height, weight),round,1)
```

문제4. 각 변수별로 unique한 값이 몇 개씩 존재하는지 파악해주세요.

```
#train %>% lapply(n_distinct)
```

문제5. 특정 변수의 경우 공백(“”)으로 처리된 값들이 존재합니다. 해당 값들을 찾아 NA로 바꿔주세요.

```
na_num<-c()
for (i in 1:ncol(train)){
  x<-which(train[,,i]=='') %>% length
  # 열 전체에 대해 TRUE/FALSE 값을 반환하기 때문에 값을 셀 수 있다.
  na_num<-c(na_num,x)}

vacancy<-data.frame(colnames(train),na_num)#vacancy %>% view
```

```
vacancy
```

```
##                  colnames.train. na_num
## 1                             age      0
## 2                             bmi      0
## 3               elective_surgery      0
## 4                      ethnicity    128
## 5                         gender      1
## 6                         height      0
## 7                         weight      0
## 8             apache_2_diagnosis      0
## 9            apache_3j_diagnosis      0
## 10         apache_post_operative      0
## 11                    arf_apache      0
## 12               gcs_eyes_apache      0
## 13              gcs_motor_apache      0
## 14             gcs_unable_apache      0
## 15             gcs_verbal_apache      0
## 16             heart_rate_apache      0
## 17              intubated_apache      0
## 18                    map_apache      0
## 19              resprate_apache      0
## 20                  temp_apache      0
## 21             ventilated_apache      0
## 22                 d1_diasbp_max      0
## 23                 d1_diasbp_min      0
## 24              d1_heartrate_max      0
## 25              d1_heartrate_min      0
## 26                   d1_mbp_max      0
## 27                   d1_mbp_min      0
## 28              d1_resprate_max      0
## 29              d1_resprate_min      0
## 30                  d1_spo2_max      0
## 31                  d1_spo2_min      0
## 32                 d1_sysbp_max      0
## 33                 d1_sysbp_min      0
## 34                  d1_temp_max      0
## 35                  d1_temp_min      0
## 36                h1_diasbp_max      0
## 37                h1_diasbp_min      0
## 38             h1_heartrate_max      0
## 39             h1_heartrate_min      0
## 40                   h1_mbp_max      0
## 41                   h1_mbp_min      0
## 42              h1_resprate_max      0
## 43              h1_resprate_min      0
## 44                  h1_spo2_max      0
## 45                  h1_spo2_min      0
## 46                 h1_sysbp_max      0
## 47                 h1_sysbp_min      0
## 48               d1_glucose_max      0
## 49               d1_glucose_min      0
## 50             d1_potassium_max      0
## 51             d1_potassium_min      0
## 52 apache_4a_hospital_death_prob      0
## 53                          aids      0
## 54                      cirrhosis      0
```

```
## 55             diabetes_mellitus      0
## 56               hepatic_failure      0
## 57             immunosuppression      0
## 58                      leukemia      0
## 59                      lymphoma      0
## 60   solid_tumor_with_metastasis      0
## 61          apache_3j_bodysystem    159
## 62           apache_2_bodysystem    159
## 63                hospital_death      0
```

```
na_num<-c()
for (i in 1:ncol(train)){
  x<-sum(train[,,i]=='')
  na_num<-c(na_num,x)
  na_num %<>%ifelse(is.na(.),0,.) }

vacancy2<-data.frame(colnames(train),na_num)
```

```
#공백이면 na 처리
train %<>% na_if('')
```

문제6. Apache_2_diagnosis, apache_3j_diagnosis, ethnicity 변수를 제거하고, 결측치를 파악하세요.

```
train %<>%select(-c(apache_2_diagnosis,apache_3j_diagnosis,ethnicity))

colSums(is.na(train))
```

```
##                          age                          bmi
##                          392                          330
##             elective_surgery                       gender
##                            0                            1
##                       height                       weight
##                          120                          265
##        apache_post_operative                  arf_apache
##                            0                           63
##              gcs_eyes_apache             gcs_motor_apache
##                          169                          169
##             gcs_unable_apache            gcs_verbal_apache
##                           85                          169
##             heart_rate_apache             intubated_apache
##                           74                           63
##                   map_apache              resprate_apache
##                           83                          111
##                  temp_apache            ventilated_apache
##                          388                           63
##                d1_diasbp_max                d1_diasbp_min
##                           13                           13
##             d1_heartrate_max             d1_heartrate_min
##                           10                           10
##                   d1_mbp_max                   d1_mbp_min
##                           16                           16
##              d1_resprate_max              d1_resprate_min
##                           33                           33
##                  d1_spo2_max                  d1_spo2_min
##                           32                           32
##                  d1_sysbp_max                 d1_sysbp_min
##                           12                           12
##                   d1_temp_max                  d1_temp_min
##                          226                          226
##                h1_diasbp_max                h1_diasbp_min
##                          317                          317
##             h1_heartrate_max             h1_heartrate_min
##                          245                          245
##                   h1_mbp_max                   h1_mbp_min
##                          401                          401
##              h1_resprate_max              h1_resprate_min
##                          400                          400
##                  h1_spo2_max                  h1_spo2_min
##                          381                          381
##                  h1_sysbp_max                 h1_sysbp_min
##                          316                          316
##               d1_glucose_max               d1_glucose_min
##                          493                          493
##             d1_potassium_max             d1_potassium_min
##                          839                          839
## apache_4a_hospital_death_prob                         aids
##                          736                           63
##                     cirrhosis            diabetes_mellitus
##                           63                           63
##              hepatic_failure             immunosuppression
##                           63                           63
##                     leukemia                     lymphoma
```

```
##                            63                            63
##   solid_tumor_with_metastasis          apache_3j_bodysystem
##                            63                           159
##           apache_2_bodysystem                hospital_death
##                           159                             0
```

문제7. 시각화하고, 결측치 비율이 10%가 넘어가는 변수는 제거해주세요

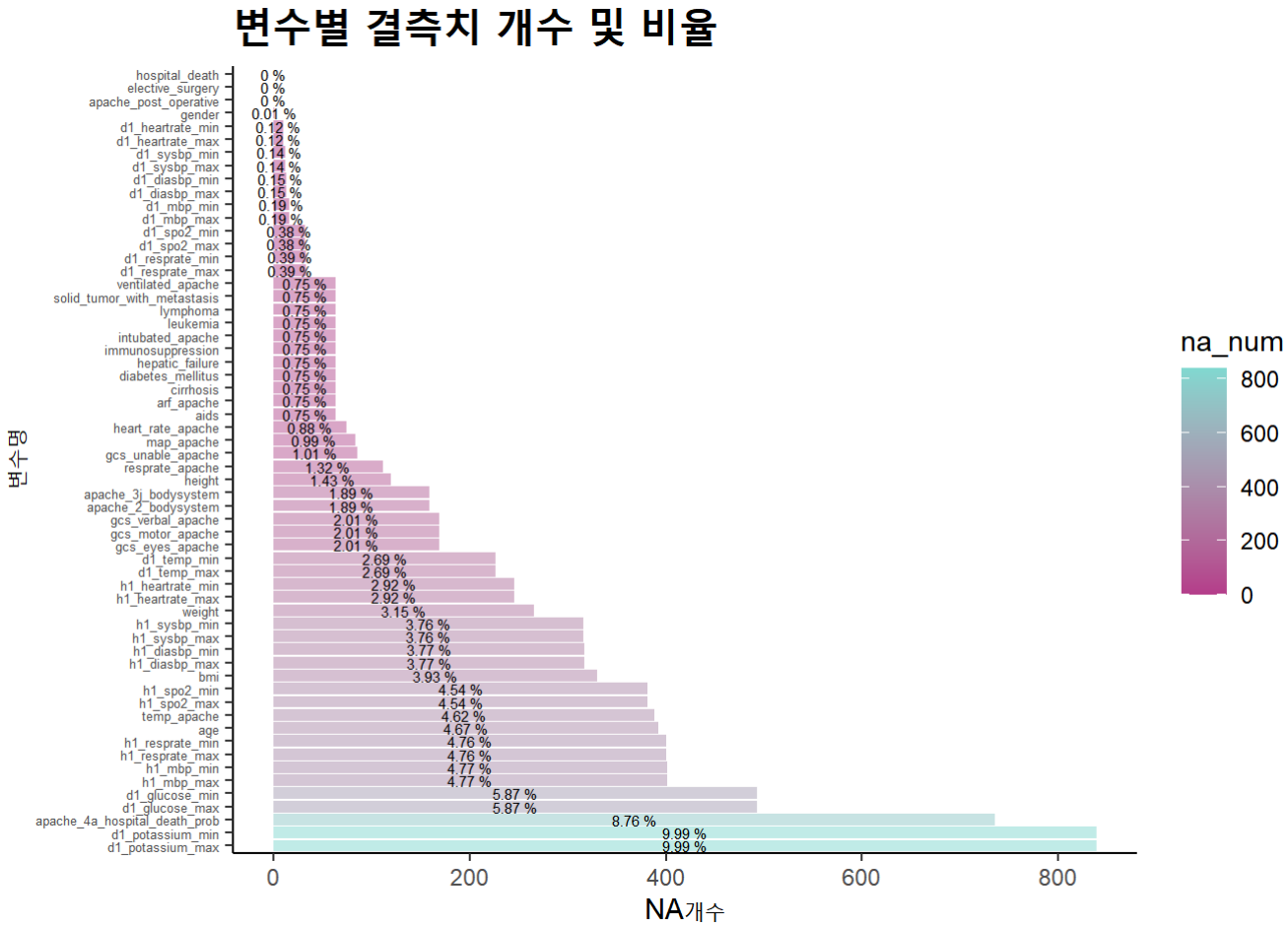
```
na<-colSums(is.na(train)) %>% as.vector
percent<-((na/nrow(train))*100)%>% round(2)
nadata<-data.frame(colnames=colnames(train),
                   percent=percent,
                   na=na,
                   percentage=paste(as.character(percent),'%'))
nadata
```

```
##                         colnames percent  na percentage
## 1                            age    4.67 392     4.67 %
## 2                            bmi    3.93 330     3.93 %
## 3               elective_surgery    0.00   0        0 %
## 4                         gender    0.01   1     0.01 %
## 5                         height    1.43 120     1.43 %
## 6                         weight    3.15 265     3.15 %
## 7         apache_post_operative    0.00   0        0 %
## 8                     arf_apache    0.75  63     0.75 %
## 9               gcs_eyes_apache    2.01 169     2.01 %
## 10              gcs_motor_apache    2.01 169     2.01 %
## 11            gcs_unable_apache    1.01  85     1.01 %
## 12            gcs_verbal_apache    2.01 169     2.01 %
## 13            heart_rate_apache    0.88  74     0.88 %
## 14             intubated_apache    0.75  63     0.75 %
## 15                   map_apache    0.99  83     0.99 %
## 16              resprate_apache    1.32 111     1.32 %
## 17                  temp_apache    4.62 388     4.62 %
## 18            ventilated_apache    0.75  63     0.75 %
## 19                d1_diasbp_max    0.15  13     0.15 %
## 20                d1_diasbp_min    0.15  13     0.15 %
## 21              d1_heartrate_max    0.12  10     0.12 %
## 22              d1_heartrate_min    0.12  10     0.12 %
## 23                    d1_mbp_max    0.19  16     0.19 %
## 24                    d1_mbp_min    0.19  16     0.19 %
## 25               d1_resprate_max    0.39  33     0.39 %
## 26               d1_resprate_min    0.39  33     0.39 %
## 27                   d1_spo2_max    0.38  32     0.38 %
## 28                   d1_spo2_min    0.38  32     0.38 %
## 29                  d1_sysbp_max    0.14  12     0.14 %
## 30                  d1_sysbp_min    0.14  12     0.14 %
## 31                    d1_temp_max    2.69 226     2.69 %
## 32                    d1_temp_min    2.69 226     2.69 %
## 33                 h1_diasbp_max    3.77 317     3.77 %
## 34                 h1_diasbp_min    3.77 317     3.77 %
## 35              h1_heartrate_max    2.92 245     2.92 %
## 36              h1_heartrate_min    2.92 245     2.92 %
## 37                    h1_mbp_max    4.77 401     4.77 %
## 38                    h1_mbp_min    4.77 401     4.77 %
## 39               h1_resprate_max    4.76 400     4.76 %
## 40               h1_resprate_min    4.76 400     4.76 %
## 41                   h1_spo2_max    4.54 381     4.54 %
## 42                   h1_spo2_min    4.54 381     4.54 %
## 43                  h1_sysbp_max    3.76 316     3.76 %
## 44                  h1_sysbp_min    3.76 316     3.76 %
## 45                d1_glucose_max    5.87 493     5.87 %
## 46                d1_glucose_min    5.87 493     5.87 %
## 47              d1_potassium_max    9.99 839     9.99 %
## 48              d1_potassium_min    9.99 839     9.99 %
## 49 apache_4a_hospital_death_prob    8.76 736     8.76 %
## 50                          aids    0.75  63     0.75 %
## 51                     cirrhosis    0.75  63     0.75 %
## 52            diabetes_mellitus    0.75  63     0.75 %
## 53               hepatic_failure    0.75  63     0.75 %
## 54            immunosuppression    0.75  63     0.75 %
```

```
## 55                      leukemia    0.75  63      0.75 %
## 56                      lymphoma    0.75  63      0.75 %
## 57   solid_tumor_with_metastasis    0.75  63      0.75 %
## 58          apache_3j_bodysystem    1.89 159      1.89 %
## 59           apache_2_bodysystem    1.89 159      1.89 %
## 60                hospital_death    0.00   0         0 %
```

```
options(repr.plot.width = 15, repr.plot.height =15)

nadata %>%
  ggplot(aes(reorder(colnames,-na),na))+
  geom_bar(aes(fill=na),stat='identity',alpha=0.5)+
  scale_fill_gradient(low='#B43C8A',high='#81D8D0')+
  labs(title='변수별 결측치 개수 및 비율',
       x='변수명',y='NA개수',fill='na_num')+
  geom_text(aes(label=percentage),size=2,
            position=position_stack(vjust=0.5))+
  theme_classic()+
  theme(plot.title=element_text(face='bold',size=18),
        axis.text.y= element_text(size=5))+
  coord_flip()
```




```
remove<-nadata %>% arrange(-percent)

train %<>% select(-(remove[c(1,2),1] %>% as.vector))
```

문제8. Hmisc 패키지를 설치하고, 변수별 결측치의 발생 패턴을 시각화하고 간단히 설명해주세요.

- Hmisc 패키지는 결측치를 시각화하는데 사용하는 패키지이며, naclus를 사용하세요.
- Method는 “average”를 사용합니다.
- 같은 형태의 Tree에 위치하면 결측치의 발생 패턴이 유사하다는 의미입니다

```
#install.packages('Hmisc')
library(Hmisc)
```

```
## 필요한 패키지를 로딩중입니다: survival
```

```
##
## 다음의 패키지를 부착합니다: 'survival'
```

```
## The following object is masked from 'package:caret':
##
##     cluster
```

```
## 필요한 패키지를 로딩중입니다: Formula
```

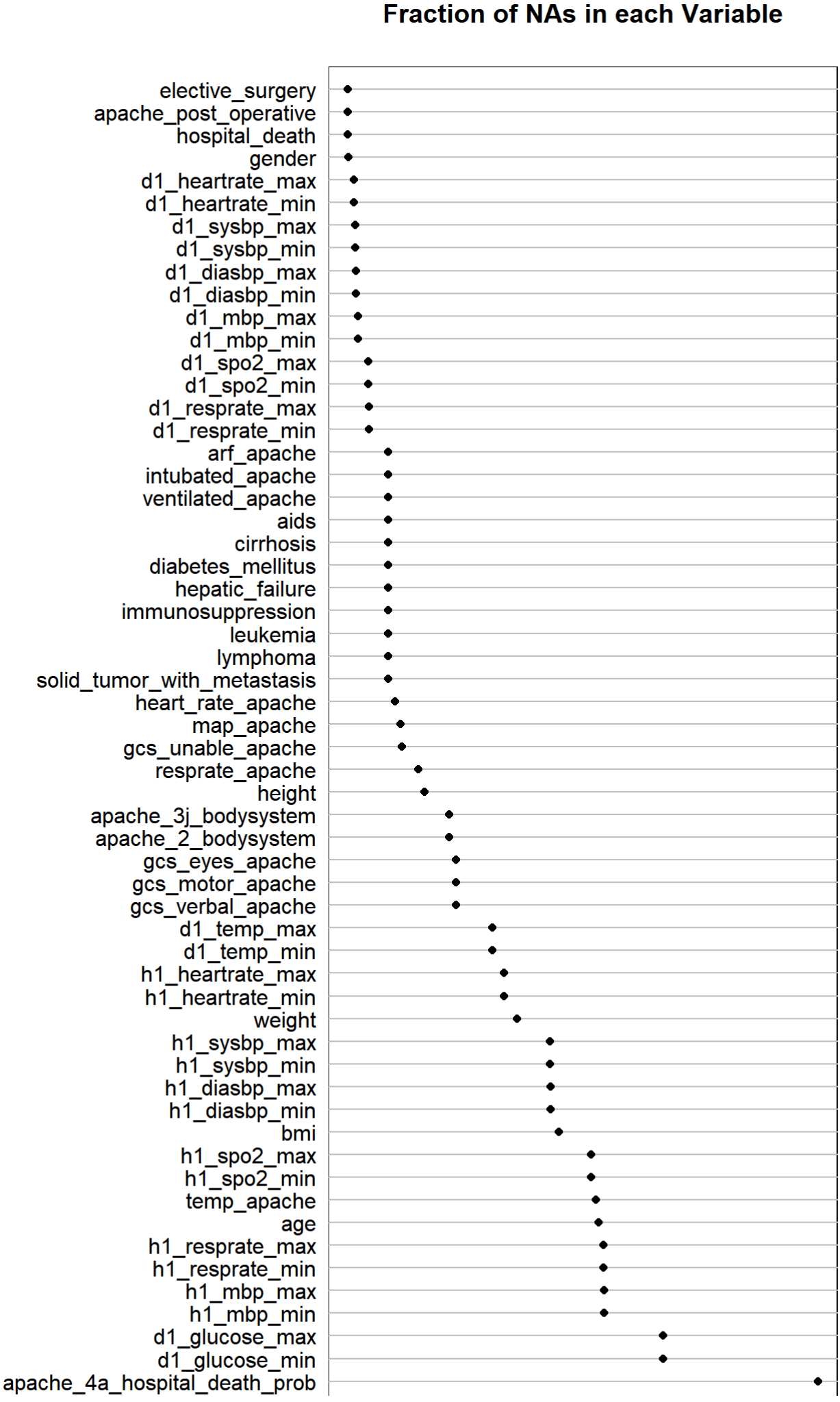
```
##
## 다음의 패키지를 부착합니다: 'Hmisc'
```

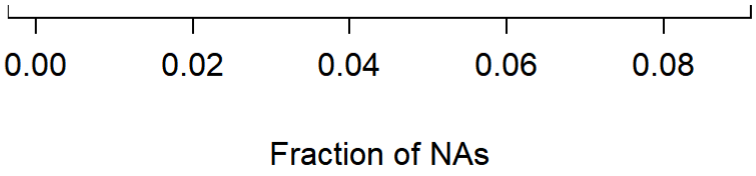
```
## The following objects are masked from 'package:dplyr':
##
##     src, summarize
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

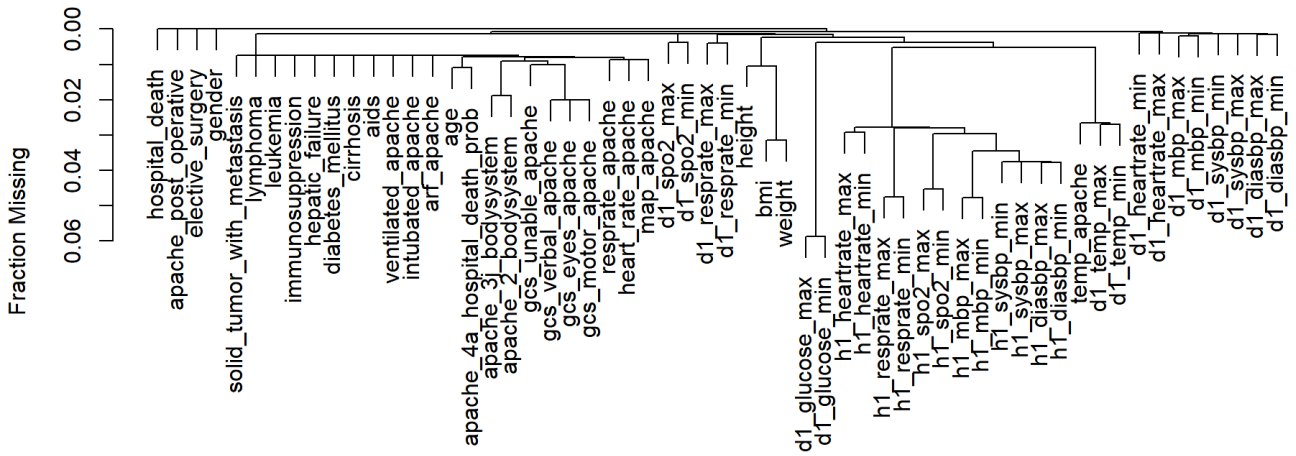
```
options(repr.plot.width = 10, repr.plot.height =30)

na_patterns <- naclus(train,method='average')
naplot(na_patterns, 'na per var')
```


Fraction of NAs in each Variable
elective_surgery
apache_post_operative
hospital_death
gender
d1_heartrate_max
d1_heartrate_min
d1_sysbp_max
d1_sysbp_min
d1_diasbp_max
d1_diasbp_min
d1_mbp_max
d1_mbp_min
d1_spo2_max
d1_spo2_min
d1_resprate_max
d1_resprate_min
arf_apache
intubated_apache
ventilated_apache
aids
cirrhosis
diabetes_mellitus
hepatic_failure
immunosuppression
leukemia
lymphoma
solid_tumor_with_metastasis
heart_rate_apache
map_apache
gcs_unable_apache
resprate_apache
height
apache_3j_bodysystem
apache_2_bodysystem
gcs_eyes_apache
gcs_motor_apache
gcs_verbal_apache
d1_temp_max
d1_temp_min
h1_heartrate_max
h1_heartrate_min
weight
h1_sysbp_max
h1_sysbp_min
h1_diasbp_max
h1_diasbp_min
bmi
h1_spo2_max
h1_spo2_min
temp_apache
age
h1_resprate_max
h1_resprate_min
h1_mbp_max
h1_mbp_min
d1_glucose_max
d1_glucose_min
apache_4a_hospital_death_prob




```
plot(na_patterns)
```



문제9. Age, bmi 변수에 결측치가 있는 경우 해당하는 row를 제거해주세요

```
train %<>%
  filter(!is.na(age)) %>%
  filter(!is.na(bmi))
```

문제10. 범주형에 해당하는 변수는 factor형으로, 그렇지 않은 변수는 수치형으로 바꿔주세요

```
train %>%
  select_if(summarise_all(.,n_distinct)<20) %>%
  lapply(unique)
```

```
## $elective_surgery
## [1] 1 0
##
## $gender
## [1] "M" "F" NA
##
## $apache_post_operative
## [1] 1 0
##
## $arf_apache
## [1] 0 1
##
## $gcs_eyes_apache
## [1]  4  1  2  3 NA
##
## $gcs_motor_apache
## [1]  6  5  4  1  3  2 NA
##
## $gcs_unable_apache
## [1]  0  1 NA
##
## $gcs_verbal_apache
## [1]  1  4  5  3  2 NA
##
## $intubated_apache
## [1] 1 0
##
## $ventilated_apache
## [1] 1 0
##
## $aids
## [1] 0 1
##
## $cirrhosis
## [1] 0 1
##
## $diabetes_mellitus
## [1] 0 1
##
## $hepatic_failure
## [1] 0 1
##
## $immunosuppression
## [1] 0 1
##
## $leukemia
## [1] 0 1
##
## $lymphoma
## [1] 0 1
##
## $solid_tumor_with_metastasis
## [1] 0 1
##
## $apache_3j_bodysystem
```

```
##  [1] "Cardiovascular"       "Neurological"         "Sepsis"
##  [4] "Gastrointestinal"     "Metabolic"            "Respiratory"
##  [7] "Hematological"        "Trauma"               NA
## [10] "Genitourinary"        "Musculoskeletal/Skin" "Gynecological"
##
## $apache_2_bodysystem
##  [1] "Undefined diagnoses" "Neurologic"          "Cardiovascular"
##  [4] "Gastrointestinal"    "Metabolic"           "Respiratory"
##  [7] "Haematologic"        "Trauma"              NA
## [10] "Renal/Genitourinary" "Undefined Diagnoses"
##
## $hospital_death
## [1] 0 1
```

```
library(DescTools)
```

```
##
## 다음의 패키지를 부착합니다: 'DescTools'
```

```
## The following objects are masked from 'package:Hmisc':
##
##     %nin%, Label, Mean, Quantile
```

```
## The following object is masked from 'package:data.table':
##
##     %like%
```

```
## The following objects are masked from 'package:caret':
##
##     MAE, RMSE
```

```
train %<>%
  mutate_if(summarise_all(.,n_distinct)<20,
            function(x){ifelse(is.na(x),Mode(x,na.rm=T),x)}) %>%
  mutate_if(summarise_all(.,n_distinct)<20,as.factor)
```

문제11. 범주형 변수에 결측치가 발생했다면 해당 변수의 최빈값(mode)으로, 수치형 변수에 발생했다면 해당 변수의 중간값으로 결측치를 대체해주세요.

- R에서는 Python과 달리 최빈값을 불러오는 함수가 존재하지 않습니다.
- 첨부된 txt 파일에 있는 최빈값 함수를 활용하셔도 됩니다.
- 많은 변수에 대해 동시에 함수를 적용해야 하므로 mutate_if 함수를 이용하면 편합니다.

```
train %<>%
  mutate_if(!(summarise_all(.,n_distinct)<20),as.numeric) %>%
  mutate_if(is.numeric,
            function(x){ifelse(is.na(x),median(x,na.rm=T),x)})
```

문제12. 결측치가 잘 대체되었는지 변수별로 결측치의 개수를 다시 확인해보세요.

```
train %>% is.na %>% colSums
```

```
##                          age                          bmi
##                            0                            0
##             elective_surgery                       gender
##                            0                            0
##                       height                       weight
##                            0                            0
##         apache_post_operative                  arf_apache
##                            0                            0
##              gcs_eyes_apache             gcs_motor_apache
##                            0                            0
##             gcs_unable_apache             gcs_verbal_apache
##                            0                            0
##             heart_rate_apache             intubated_apache
##                            0                            0
##                   map_apache              resprate_apache
##                            0                            0
##                  temp_apache             ventilated_apache
##                            0                            0
##                d1_diasbp_max                d1_diasbp_min
##                            0                            0
##             d1_heartrate_max             d1_heartrate_min
##                            0                            0
##                   d1_mbp_max                   d1_mbp_min
##                            0                            0
##              d1_resprate_max              d1_resprate_min
##                            0                            0
##                  d1_spo2_max                  d1_spo2_min
##                            0                            0
##                 d1_sysbp_max                 d1_sysbp_min
##                            0                            0
##                  d1_temp_max                  d1_temp_min
##                            0                            0
##                h1_diasbp_max                h1_diasbp_min
##                            0                            0
##             h1_heartrate_max             h1_heartrate_min
##                            0                            0
##                   h1_mbp_max                   h1_mbp_min
##                            0                            0
##              h1_resprate_max              h1_resprate_min
##                            0                            0
##                  h1_spo2_max                  h1_spo2_min
##                            0                            0
##                 h1_sysbp_max                 h1_sysbp_min
##                            0                            0
##               d1_glucose_max               d1_glucose_min
##                            0                            0
## apache_4a_hospital_death_prob                         aids
##                            0                            0
##                     cirrhosis            diabetes_mellitus
##                            0                            0
##               hepatic_failure            immunosuppression
##                            0                            0
##                      leukemia                     lymphoma
##                            0                            0
##   solid_tumor_with_metastasis         apache_3j_bodysystem
```

```
##                             0                             0
##          apache_2_bodysystem                hospital_death
##                             0                             0
```

문제13. Test set에 대해서도 Train set과 동일한 전처리를 진행해주세요.

```
test %<>%
  select(-contains(c('_id','icu_','noninvasive_'))) %>%
  select(-V84) %>%
  mutate_at(vars(bmi, height, weight),round,1) %>%
  select(-c(apache_2_diagnosis,apache_3j_diagnosis,ethnicity)) %>%
  select(-(remove[c(1,2),1] %>% as.vector)) %>%
  na_if('') %>%
  filter(!is.na(age)) %>%
  filter(!is.na(bmi)) %>%
  mutate_if(summarise_all(train,n_distinct)<20,
            function(x){ifelse(is.na(x),Mode(x,na.rm=T),x)}) %>%
  mutate_if(summarise_all(train,n_distinct)<20,as.factor) %>%
  mutate_if(!(summarise_all(train,n_distinct)<20),as.numeric) %>%
  mutate_if(is.numeric,
            function(x){ifelse(is.na(x),median(x,na.rm=T),x)})
```

Chapter 2 : RandomForest & Grid Search

문제1. Train 데이터를 train과 valid set으로 분리하세요. (7:3의 비율로 구성, seed : 2930)

```
set.seed(2930)

index<-createDataPartition(train$age,p=0.7,list=F)
train_rf<-train[index,]
valid_rf<-train[-index,]
```

문제3. 사용자가 임의로 파라미터의 조합을 지정하는 Grid-Search 알고리즘 방법을 적용하기 위해서, 데이터프레임을 만들어주세요.

```
grid<-expand.grid(mtry=c(6,8,10,12),ntree=c(300,400,500),F1_score=NA)
```

문제4. randomForest, MLmetrics 패키지를 설치하고 불러와주세요.

```
library(randomForest)
```

```
## randomForest 4.7-1.1
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## 다음의 패키지를 부착합니다: 'randomForest'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
library(MLmetrics)
```

```
##
## 다음의 패키지를 부착합니다: 'MLmetrics'
```

```
## The following objects are masked from 'package:DescTools':
##
##     AUC, Gini, MAE, MAPE, MSE, RMSE
```

```
## The following objects are masked from 'package:caret':
##
##     MAE, RMSE
```

```
## The following object is masked from 'package:base':
##
##     Recall
```

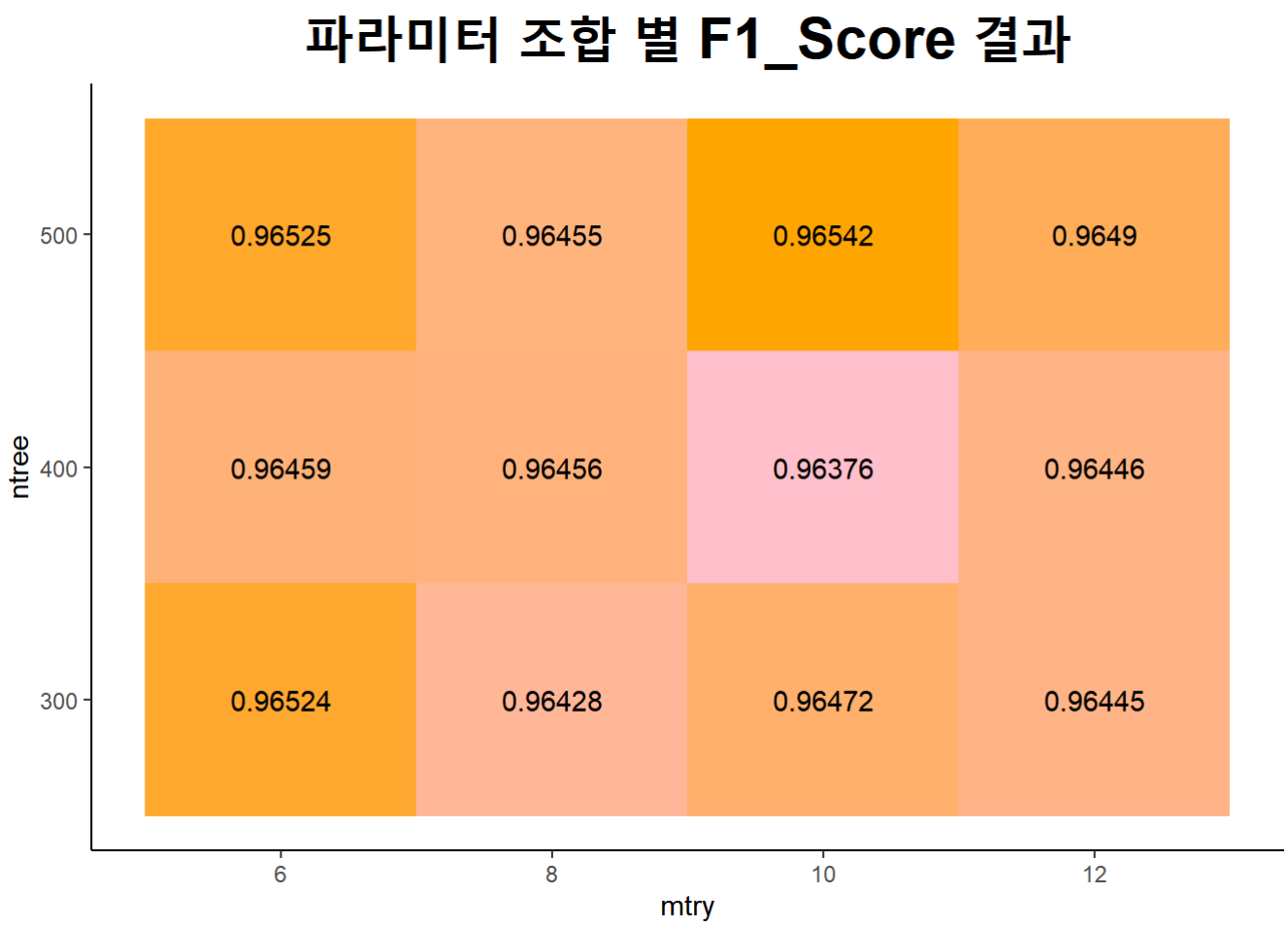
```
for(i in 1:nrow(grid)){
  model<-randomForest(hospital_death~.,
                      data=train_rf,
                      mtry=grid$mtry[i],
                      ntree=grid$ntree[i],
                      importance=TRUE)

  pred=predict(model,newdata=valid_rf,type='response')
  grid$F1_score[i]<-F1_Score(pred,valid_rf$hospital_death)
}
```

```
grid
```

```
##    mtry ntree  F1_score
## 1     6   300 0.9652352
## 2     8   300 0.9642776
## 3    10   300 0.9647166
## 4    12   300 0.9644485
## 5     6   400 0.9645937
## 6     8   400 0.9645616
## 7    10   400 0.9637566
## 8    12   400 0.9644647
## 9     6   500 0.9652510
## 10    8   500 0.9645455
## 11   10   500 0.9654231
## 12   12   500 0.9649043
```

```
grid %>% ggplot(aes(x=mtry,y=ntree,fill=F1_score))+
  geom_tile()+
  scale_fill_gradient(low='pink',high='orange')+
  geom_text(aes(label=F1_score %>% round(5) %>% as.character))+
  theme_classic()+
  ggtitle("파라미터 조합 별 F1_Score 결과")+
  theme(plot.title=element_text(size=22,hjust=0.5,face='bold'),
        legend.position='none')
```

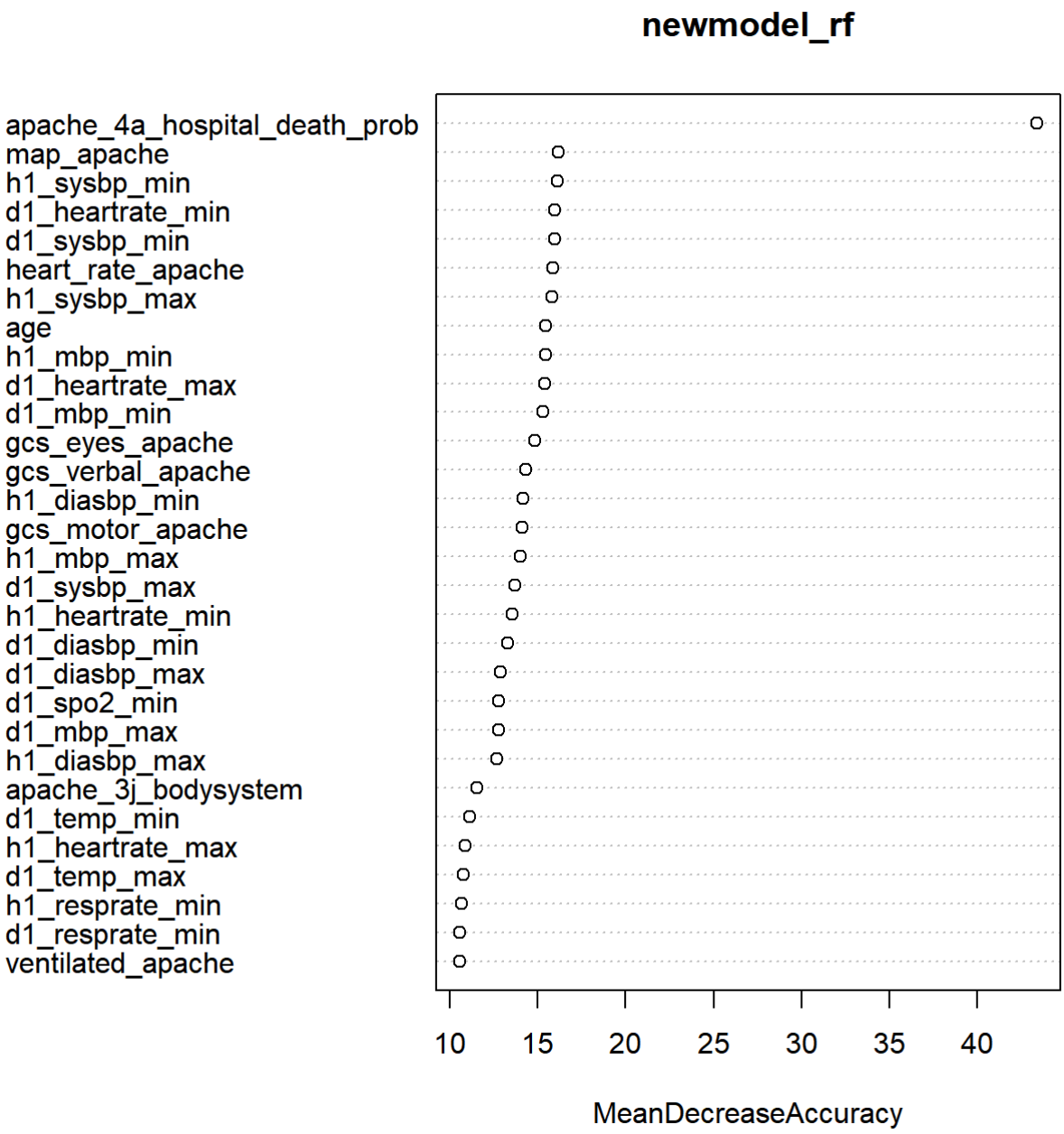


가장 F1_score가 높은 파라미터 조합을 찾고, 해당 결과로 모델을 다시 학습시키세요.

```
max<-which(grid$F1_score==max(grid$F1_score))
```

```
newmodel_rf<-randomForest(hospital_death~.,
                          data=train,
                          ntree=grid$ntree[max],
                          mtry=grid$mtry[max],
                          importance=TRUE)
```

```
varImpPlot(newmodel_rf,type=1)
```



Chapter 3 : XGBoost & Random Tuning

문제1. Train 데이터를 train과 valid set으로 분리하세요. (7:3의 비율로 구성, seed : 2930)

```
set.seed(2930)

index<-createDataPartition(train$age,p=0.7,list=F)
train_xgb<-train[index,]
valid_xgb<-train[-index,]
```

문제2. XGBoost를 사용하기 위해, train과 valid set에 존재한 모든 범주형 변수에 대해서 One-hot Encoding을 실행해주세요.

- XGBoost는 범주형 자료를 인식하지 못하므로 숫자로 인식시켜야 합니다
- 예측해야 할 hospital_death 변수를 제외하고 dummyVars로 인코딩을 해준 뒤, 다시 합쳐주는 형식으로 진행하셔야 오류를 피할 수 있습니다.

```
target<-train %>% select(hospital_death)
xdata<-train %>% select(-hospital_death)
```

```
dmy <- dummyVars(~., data=xdata)
train_new <- data.frame(predict(dmy, newdata=xdata))
```

```
train_new<-cbind(train_new,target)
```

문제4. 하이퍼파라미터의 범위를 아래와 같이 설정하고, F1_score가 NA로 비워진 데이터프레임을 만드세요.

- 랜덤튜닝은 총 10번 진행합니다.
- max_depth = 4 ~ 10 / min_child_weight = 4 ~ 10 / subsample = 0.5 ~ 1 / colsample_bytree = 0.5 ~ 1

```
set.seed(2930)
rdtn<-data.frame(iterations=c(1:10),
                 max_depth=sample(x=4:10,size=10,replace=T),
                 min_child_weight=sample(x=4:10,size=10,replace=T),
                 subsample=runif(10,0.5,1),
                 colsample_bytree=runif(10,0.5,1),
                 F1_score=NA)
```

문제5. xgboost 패키지를 설치하고, 패키지를 불러와주세요

```
library(xgboost)
```

```
##
## 다음의 패키지를 부착합니다: 'xgboost'
```

```
## The following object is masked from 'package:dplyr':
##
##     slice
```

```
#index 만들어주기
folds<-createFolds(train_new,k=5)
```

문제7. 5-fold CV를 XGBoost에 적용시켜보겠습니다. 파라미터 조합을 변경시키는 첫번째 for문과 valid set을 지속적으로 변경시키는 두번째 for문으로 구성됩니다.

```
set.seed(2930)

f1score<-c()
x<-train_new%>% select(-hospital_death) %>% as.matrix
y<-train_new%>% select(hospital_death) %>% as.matrix %>% as.numeric

for(i in 1:nrow(rdtn)){
  param<-list(max_depth=rdtn$max_depth[i],
              min_child_weight=rdtn$min_child_weight[i],
              subsample=rdtn$subsample[i],
              colsample_bytree=rdtn$colsample_bytree[i])

  for(j in c(1:5)){

    index<-folds[[j]]

    train_x<-x[-index,]
    train_y<-y[-index]

    valid_x<-x[index,]
    valid_y<-y[index]

    cv_train<-xgb.DMatrix(train_x,label = train_y)
    cv_val<-xgb.DMatrix(valid_x,label = valid_y)

    xgb<-xgb.train(params=param,data=cv_train,
                   eta=0.01,nrounds=100)

    pred<-predict(xgb,cv_val)
    f1score<-c(f1score,F1_Score(valid_y,round(pred)))
  }
  rdtn$F1_score[i]<-mean(f1score)
}
```

문제8. 가장 좋은 파라미터 조합을 출력하고, 해당 파라미터 조합으로 모델을 다시 학습시키세요.

```
max<-which(rdtn$F1_score==max(rdtn$F1_score))
max<-max[1]
rdtn[1,]
```

```
##   iterations max_depth min_child_weight subsample colsample_bytree  F1_score
## 1          1         9                7 0.5704327        0.8031388 0.9388085
```

```
xgbtrain<-xgb.DMatrix(x,label=y)
```

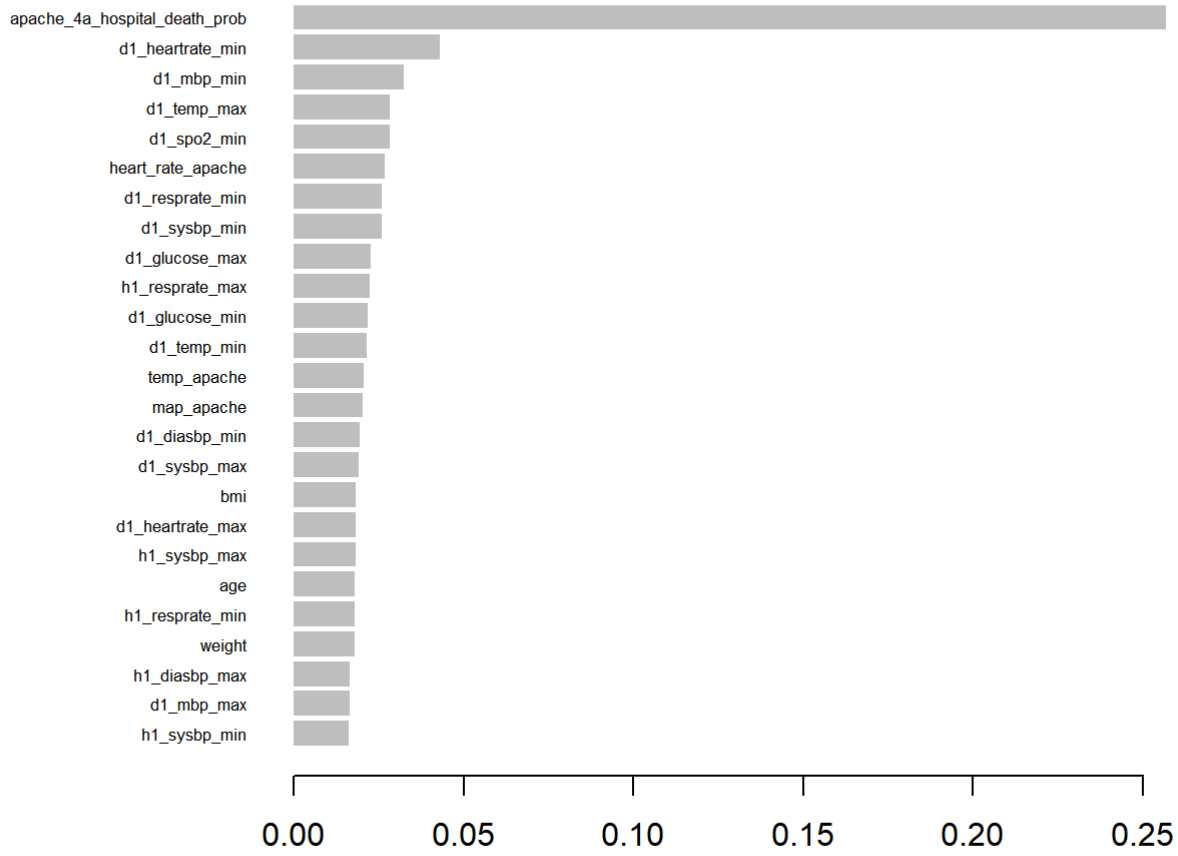
```
newparam<-list(max_depth=rdtn$max_depth[max],
               min_child_weight=rdtn$min_child_weight[max],
               subsample=rdtn$subsample[max],
               colsample_bytree=rdtn$colsample_bytree[max])

newmodel_xg<-xgb.train(params=newparam,data=xgbtrain,
                   eta=0.01,nrounds=100)
```

문제9. 변수의 중요도를 Importance matrix로 시각화하고, 어떤 변수가 환자의 사망 여부에 영향을 미치는지 서술해주세요.

```
importance_matrix = xgb.importance(colnames(xgbtrain), model = newmodel_xg)
```

```
xgb.plot.importance(importance_matrix[1:25])
```



Chapter 4 : RandomForest vs XGBoost

문제1. RandomForest의 최고 파라미터 조합을 사용하여 모델을 학습시킨 후, test 데이터에 대해 F1-Score를 계산해보세요.

```
rf_predict<-predict(newmodel_rf,newdata=test,type='response')
```

```
F1_Score(rf_predict,test$hospital_death)
```

```
## [1] 0.9657578
```

문제2. XGboost의 최고 파라미터 조합을 사용하여 모델 학습 후, test에 대해 F1-Score를 계산해보세요.

```
target<-test %>% select(hospital_death)
xdata<-test %>% select(-hospital_death)

dmy <- dummyVars(~., data=xdata)
test_new <- data.frame(predict(dmy, newdata=xdata))
test_new<-cbind(test_new,target)
```

```
test_x<-test_new %>% select(-hospital_death) %>%as.matrix #x_data set
test_y<-test_new %>% select(hospital_death) %>% as.matrix %>% as.numeric #label data

xgbtest<-xgb.DMatrix(test_x,label = test_y)
pred<-predict(newmodel_xg,newdata=xgbtest)
```

```
F1_Score(round(pred),test_y)
```

```
## [1] 0.9657252
```

```
data.frame(f1score_of_RF=F1_Score(rf_predict,test$hospital_death),
           f1score_of_XG=F1_Score(round(pred),test_y))
```

```
##   f1score_of_RF f1score_of_XG
## 1     0.9657578     0.9657252
```