midterm_32180331

32180331_김가영

2020 10 29

1. diagnosis.csv 데이터셋 소개

이 보고서에서 데이터 분석을 할 대상 자료는 diagnosis.csv로, 국민건강보험공단에서 제공하는 2016~2017년 안과 진료 데이터이다. diagnosis.csv는 총, 1,652,504 건의 진료기록을 담고 있고, 각 진료기록은 총 19항목 정보를 가지고 있는데 12항목을 선택하여 분석하고자 한다.

(1) STND_Y : 기준년도

(2) IDV_ID : 가입자 일렬번호

(3) KEY_SEQ : 진료내역 일렬번호

(4) SEX : 대상자 성별

(5) AGE_GROUP : 수진자 나이 그룹 코드(0~84세 – 5세 단위, 85세 이상 – 85+, 총 18개 그룹)

(6) SIDO : 수진자 거주지의 시도코드

(7) RECU_FR_DT : 요양 개시 일자

(8) FORM_CD : 진료형태 구분코드

(9) DSBJT_CD : 진료 과목 코드 (안과 = 12)

(10) MAIN_SICK : 주상병 분류코드

(11) SUB_SICK : 부상병 분류코드

(12) VSCN : 요양 일수

(13) RECN : 입내원 일수

(14) EDEC_ADD_RT : 심결 가산률, 요양기관 종별에 따라 가산 적용되는 진료비의 가산율

(15) EDEC_TRAMT : 심결요양급여비용총액 (=본인부담금 + 보험자부담금)

(16) EDEC_SBRDN_AMT : 본인부담금

(17) EDEC_JBRDN_AMT : 보험자부담금

(18) TOT_PRES_DD_CNT : 총 처방일수

(19) GEN_MAIN_SICK : MAIN_SICK의 상위 분류코드(앞의 3글자)만 추출하여 생성한 항목

2. 탐색적 데이터 분석 과정

2.1. 분석 대상 데이터셋 준비

먼저 분석 대상이 되는 데이터셋을 준비하기 위해 diagnosis.csv 데이터 소스를 불러왔다. setwd()를 이용하여 작업 디렉토리를 지정하고, read.csv 내장 함수를 사용하여 diagnosis.csv 데이터를 불러왔다.

열의 이름이 지정되어 있지 않아 colnames() 함수를 이용하여 열의 이름을 지정했다.

```
colnames(diagnosis) <- c("STND_Y","IDV_ID","KEY_SEQ","SEX","AGE_GROUP","SIDO","RECU_FR_DT","FORM_CD","DSBJT_CD","MAIN_SICK","SUB_SICK","VSCN","RECN","EDEC_ADD_RT","EDEC_TRAMT","EDEC_SBRDN_AMT","EDEC_JBRDN_AMT","TOT_PRES_DD_CNT","DATA_STD_DT")
colnames(diagnosis)
```

```
##  [1] "STND_Y"          "IDV_ID"          "KEY_SEQ"         "SEX"
##  [5] "AGE_GROUP"       "SIDO"            "RECU_FR_DT"      "FORM_CD"
##  [9] "DSBJT_CD"        "MAIN_SICK"       "SUB_SICK"        "VSCN"
## [13] "RECN"            "EDEC_ADD_RT"     "EDEC_TRAMT"      "EDEC_SBRDN_AMT"
## [17] "EDEC_JBRDN_AMT"  "TOT_PRES_DD_CNT" "DATA_STD_DT"
```

2.2. 데이터셋의 형태와 기본적인 내용 파악

```
str(diagnosis)
```

```
## 'data.frame':    1652503 obs. of  19 variables:
##  $ STND_Y         : int  2016 2016 2016 2016 2016 2016 2016 2016 2016 2016 ...
##  $ IDV_ID         : int  5 7 7 7 10 14 14 14 14 16 ...
##  $ KEY_SEQ        : int  21739592 43400581 37558277 8400884 54462644 47218014 40081350 45929
717 60361567 181714 ...
##  $ SEX            : int  1 2 2 2 2 2 2 2 2 2 ...
##  $ AGE_GROUP      : int  2 14 14 14 15 2 2 2 2 7 ...
##  $ SIDO           : int  41 43 43 43 41 42 42 42 42 41 ...
##  $ RECU_FR_DT     : int  20160608 20160713 20161230 20160217 20161007 20161006 20160822 2016
1026 20161108 20160430 ...
##  $ FORM_CD        : int  3 3 3 3 3 3 3 3 3 3 ...
##  $ DSBJT_CD       : int  12 12 12 12 12 12 12 12 12 12 ...
##  $ MAIN_SICK      : chr  "H521" "H2512" "H2512" "H2512" ...
##  $ SUB_SICK       : chr  "H5221" "H019" "H019" "H019" ...
##  $ VSCN           : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ RECN           : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ EDEC_ADD_RT    : num  0.15 0.15 0.15 0.15 0.15 0.15 0.15 0.15 0.15 0.15 ...
##  $ EDEC_TRAMT     : int  29380 16640 16640 18840 66670 16640 26680 28880 24770 20210 ...
##  $ EDEC_SBRDN_AMT : int  8800 4900 4900 5600 20000 4900 8000 8600 7400 6000 ...
##  $ EDEC_JBRDN_AMT : int  20580 11740 11740 13240 46670 11740 18680 20280 17370 14210 ...
##  $ TOT_PRES_DD_CNT: int  0 1 1 1 7 1 1 1 1 1 ...
##  $ DATA_STD_DT    : int  20171218 20171218 20171218 20171218 20171218 20171218 20171218 2017
1218 20171218 20171218 ...
```

```
head(diagnosis)
```

```
##   STND_Y IDV_ID  KEY_SEQ SEX AGE_GROUP SIDO RECU_FR_DT FORM_CD DSBJT_CD
## 1   2016      5 21739592   1         2   41   20160608       3       12
## 2   2016      7 43400581   2        14   43   20160713       3       12
## 3   2016      7 37558277   2        14   43   20161230       3       12
## 4   2016      7  8400884   2        14   43   20160217       3       12
## 5   2016     10 54462644   2        15   41   20161007       3       12
## 6   2016     14 47218014   2         2   42   20161006       3       12
##   MAIN_SICK SUB_SICK VSCN RECN EDEC_ADD_RT EDEC_TRAMT EDEC_SBRDN_AMT
## 1      H521    H5221    1    1        0.15      29380           8800
## 2     H2512     H019    1    1        0.15      16640           4900
## 3     H2512     H019    1    1        0.15      16640           4900
## 4     H2512     H019    1    1        0.15      18840           5600
## 5      H811     G470    1    1        0.15      66670          20000
## 6      H101     H521    1    1        0.15      16640           4900
##   EDEC_JBRDN_AMT TOT_PRES_DD_CNT DATA_STD_DT
## 1          20580              0    20171218
## 2          11740              1    20171218
## 3          11740              1    20171218
## 4          13240              1    20171218
## 5          46670              7    20171218
## 6          11740              1    20171218
```

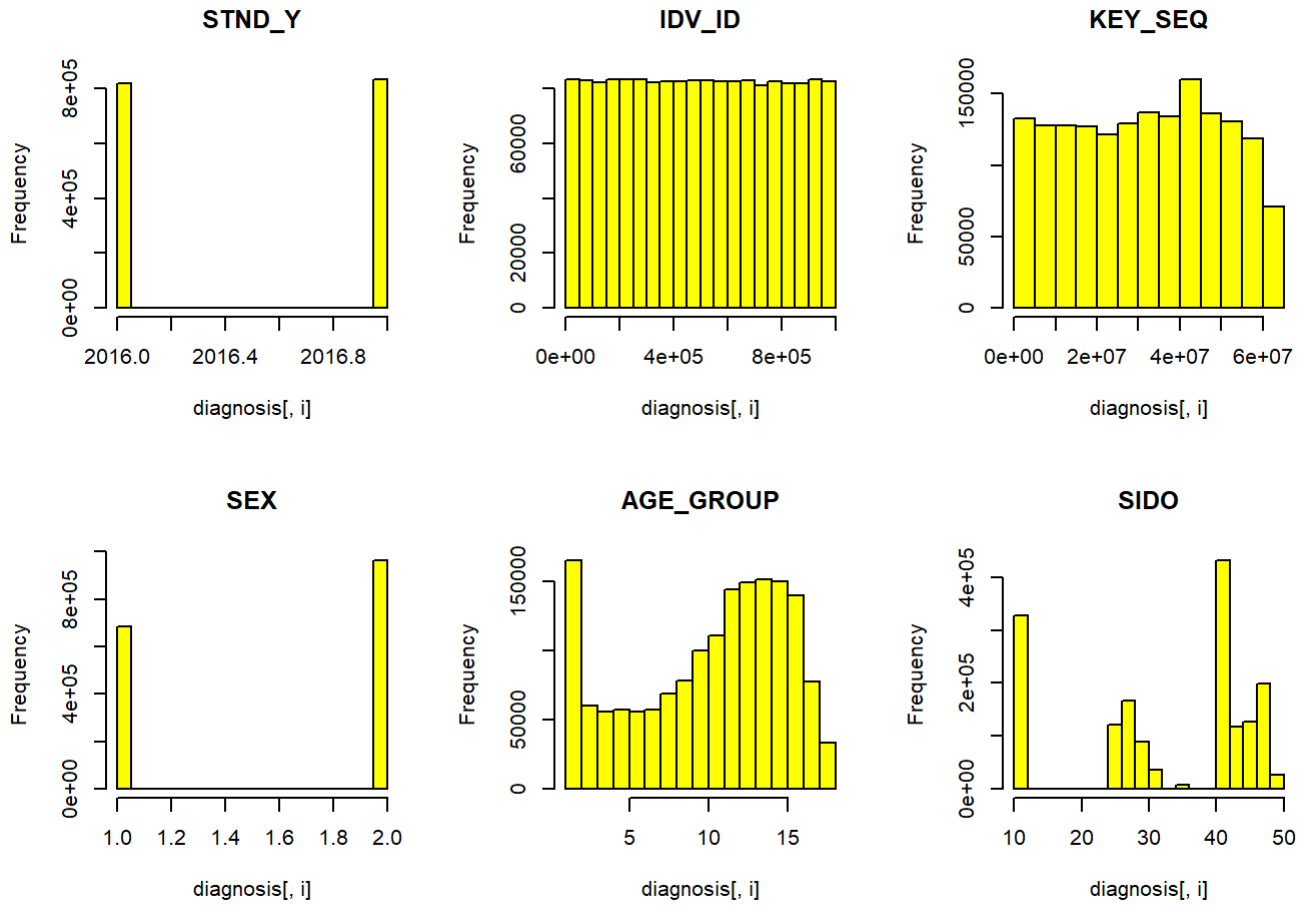
·str() 함수의 결과 : myds 데이터셋은 데이터프레임. 1652504개에 대한 정보를 담고 있으며 19개의 변수로 구성되어 있음. 19개의 변수 중 EDEC_ADD_RT 는 숫자 타입, 나머지는 정수,문자타입의 변수임.

·head() 함수의 결과 : 데이터가 저장되어 있는 형태를 확인가능.

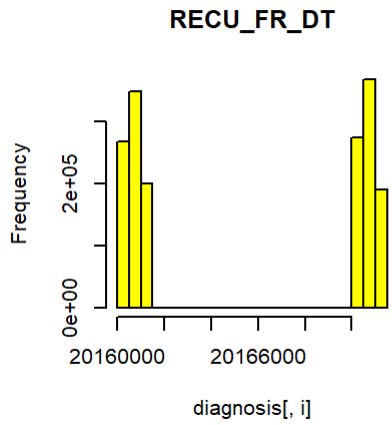
2.3. 히스토그램에 의한 관측값의 분포 확인

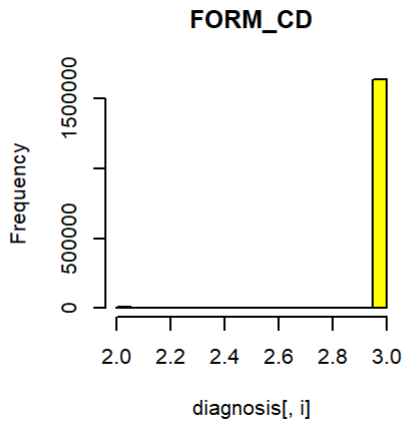
MAIN_SICK, SUB_SICK은 문자형이므로 제외하고 17개 변수에 대해 관측값들의 분포를 확인해보았다.

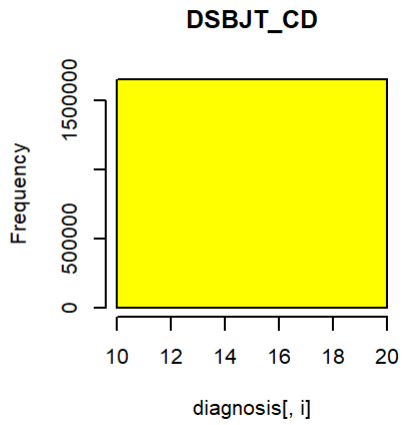
```
# 2 x 3 가상화면 분할
par(mfrow=c(2,3))
for(i in 1:9){
  hist(diagnosis[,i], main=colnames(diagnosis)[i], col="yellow")
}
```

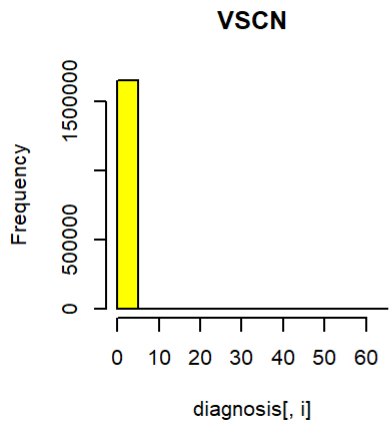


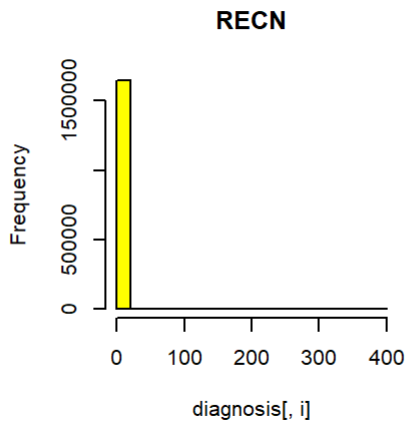

```
for(i in 12:19){
  hist(diagnosis[,i], main=colnames(diagnosis)[i], col="yellow")
}
```

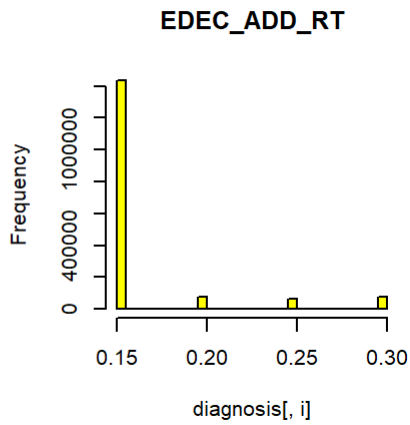




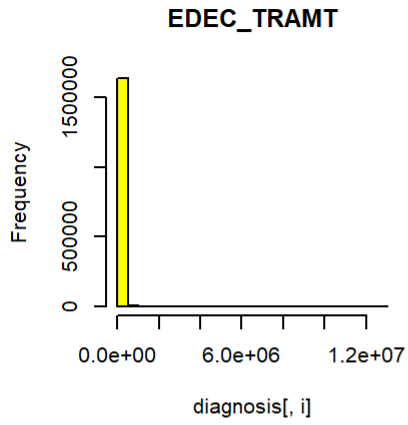


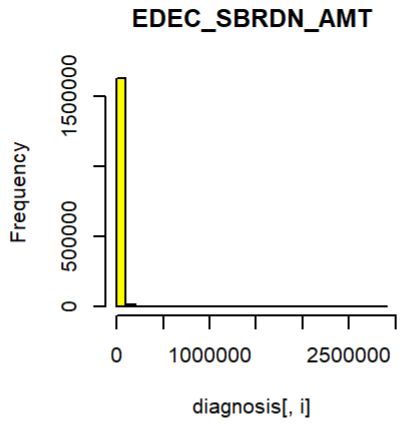


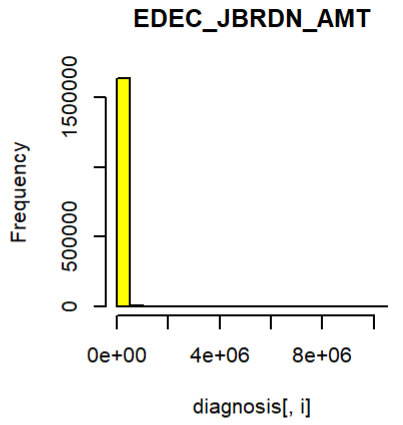


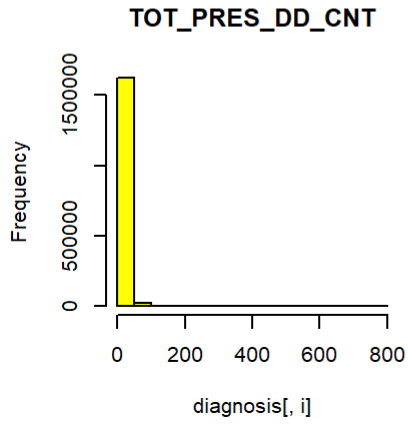


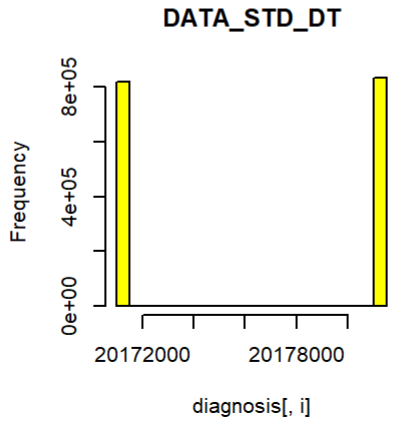

```
# 2 x 3 가상화면 분할 해제
par(mfrow=c(1,1))
```











·STND_Y : 2016, 2017년 절반씩 조사되었다는 것을 알 수 있음.

·IDV_ID, KEY_SEQ : 일련번호이므로 고르게 퍼져있음.

·SEX : 1번인 남자보다 2번인 여자가 더 많이 분포한다는 것을 알 수 있음.

·AGE_GROUP : 그룹 1(0~4세)이 가장 높게 분포하고, 그룹 $11_{15(50}$74세)도 높은 편으로 분포함.

·SIDO : 중간 중간에 관측값이 없는 빈 구간이 존재하는 특징을 보이고, 수도권인 그룹 11(서울), 그룹 41(경기도)가 높게 분포함.

·RECU_FR_DT : 2016년 초와 2017년 말에 모여서 분포함.

·FROM_CD : 3번인 의과 외래에 몰려있음.

·DSBJT_CD : 12번 안과로 모두 같아서 위와 같이 그려짐.

·VSCN : 모두 1로 한 쪽에 쏠려서 분포함.

·RECN : 모두 1로 한 쪽에 쏠려서 분포함.

·EDEC_ADD_RT : 대부분 0.15로 한 쪽에 쏠려서 분포함.

·EDEC_TRAMT, EDEC_SBRON_AMT, EDEC_JBROM_AMT : 한 쪽에 쏠려서 분포함.

·TOT_PRES_DD_CNT : 한 쪽에 몰려서 분포함.

·DATA_STD_DT : 두 개의 값으로 나뉘어 분포함.

2.4. 테이블에 의한 관측값의 분포 확인

SEX(성별)을 테이블로 나타내보았음.

```
table(diagnosis$SEX)
```

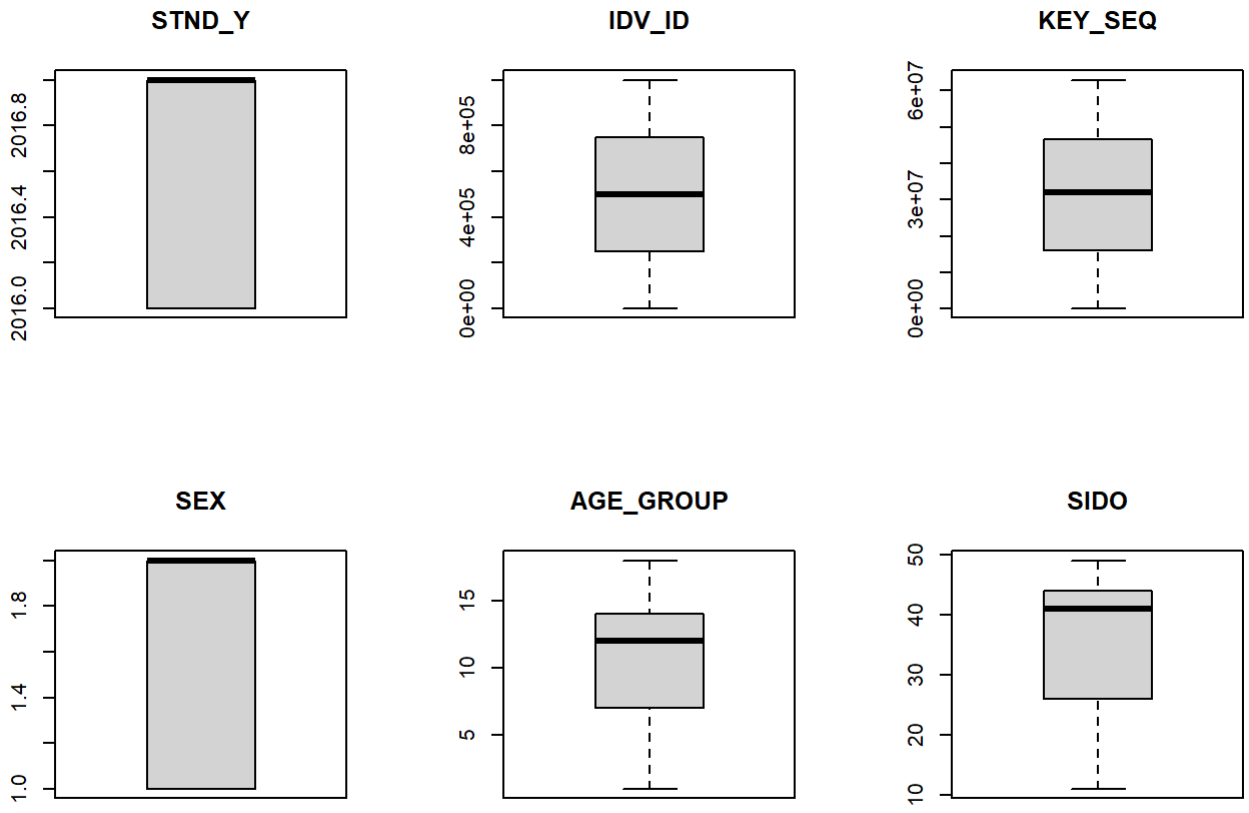
```
##
##      1      2
## 686863 965640
```

· 2.3.의 히스토그램에서도 보았듯이 2번인 여자가 약 280000명 정도 더 많이 분포한다는 것을 알 수 있음 -> 여자가 남자보다 눈에 질환이 잘 생기는 것이 아닐까라는 추측을 해볼 수 있음.

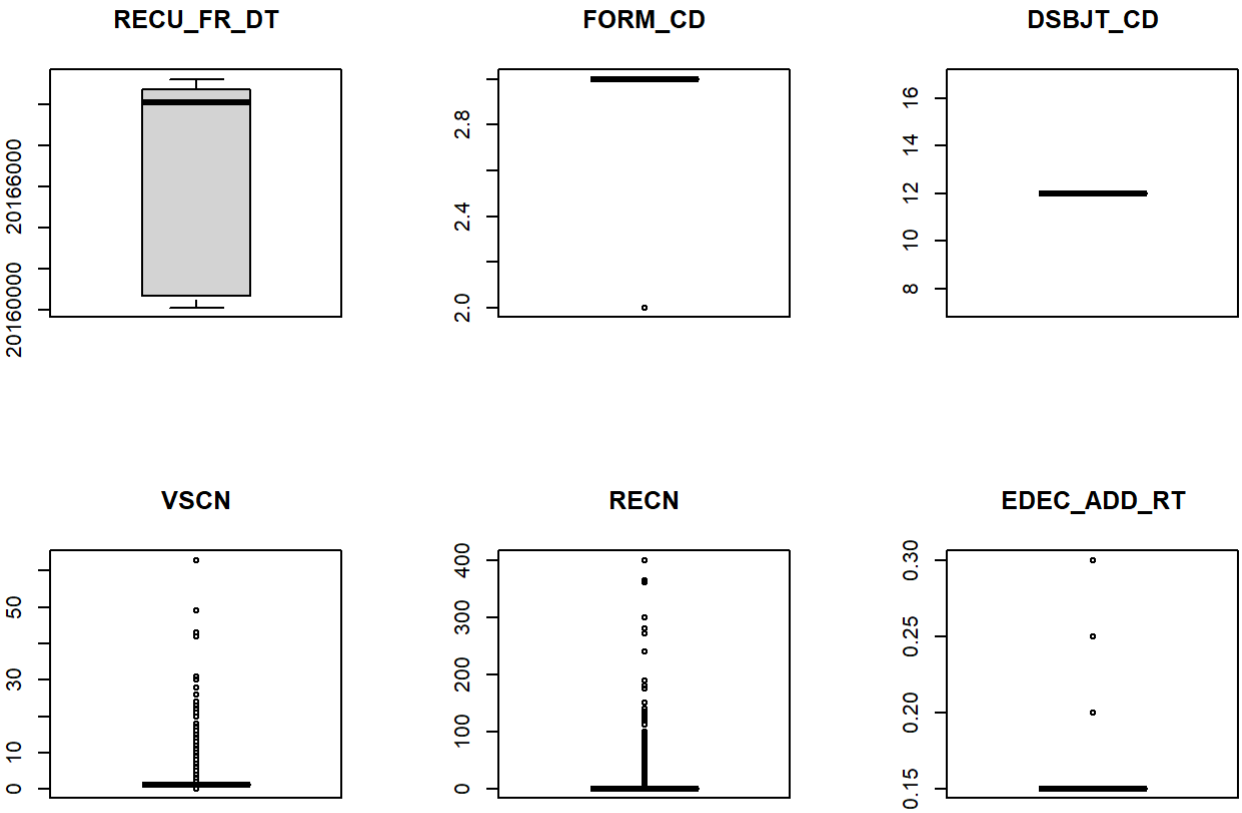
2.5. 상자그림에 의한 관측값의 분포 확인

문자형 변수를 제외한 나머지 변수에 대해 boxplot()함수를 작성하여 분포를 확인해보았다.

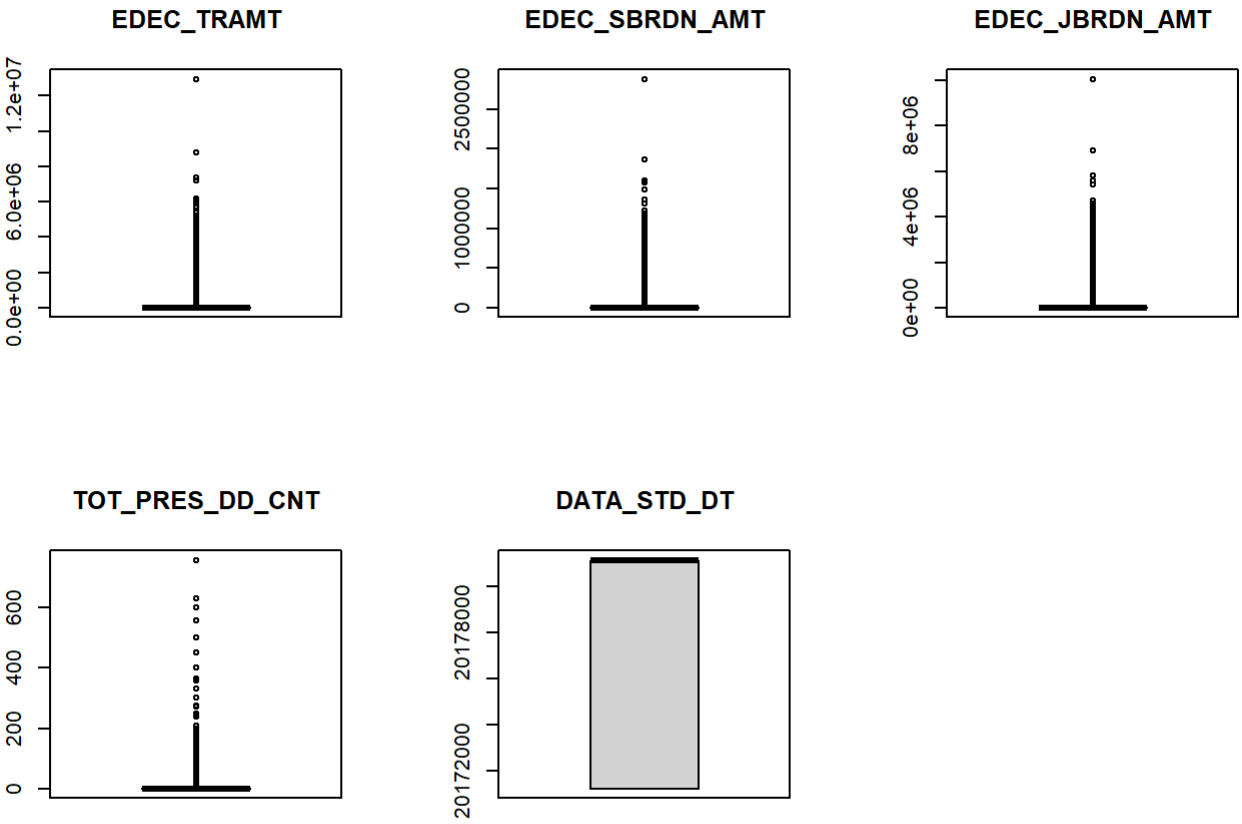
```
# 2 x 3 가상화면 분할
par(mfrow=c(2,3))
for(i in 1:9){
  boxplot(diagnosis[,i], main=colnames(diagnosis)[i])
}
```




```
for(i in 12:19){
  boxplot(diagnosis[,i], main=colnames(diagnosis)[i])
}
```




```
# 2 x 3 가상화면 분할 해제
par(mfrow=c(1,1))
```



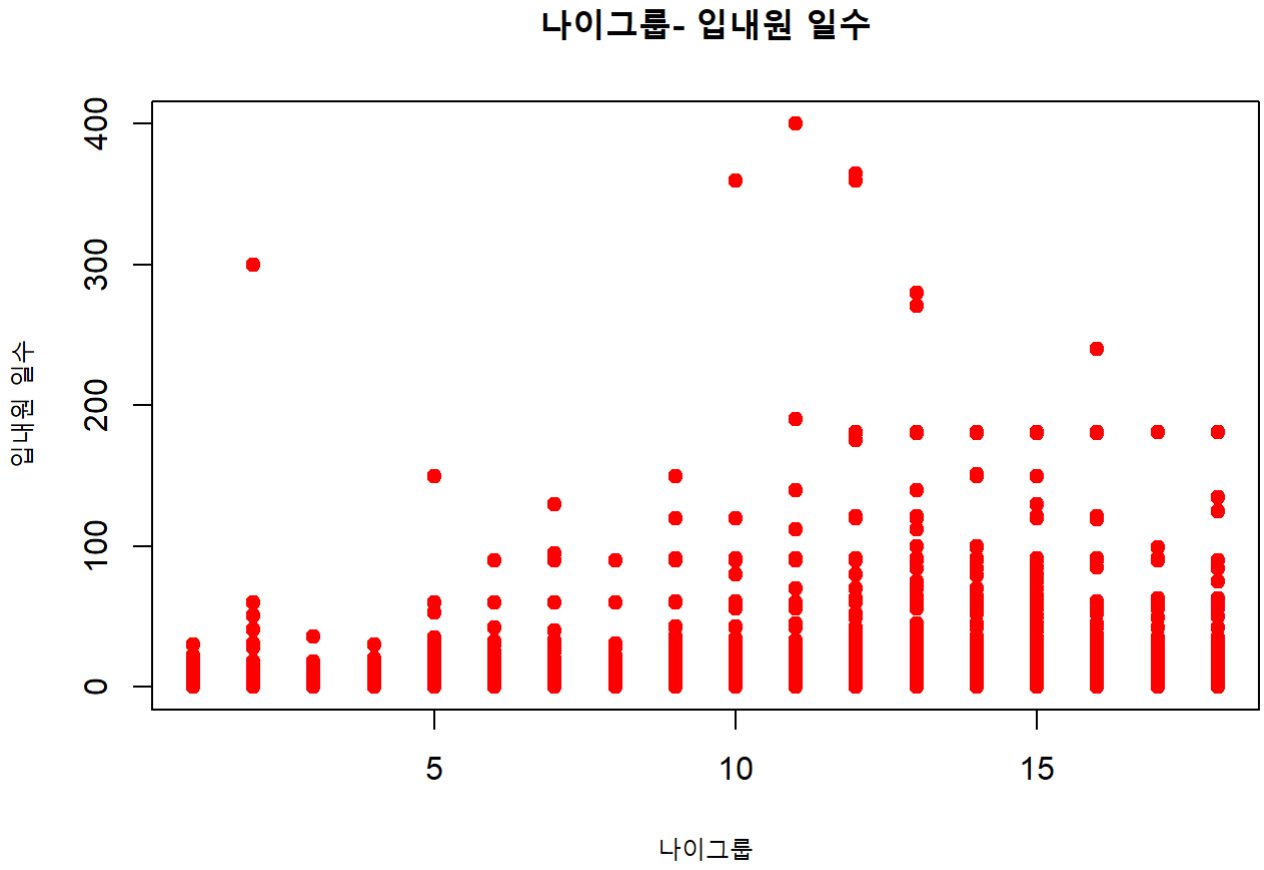
· STND_Y,IDV_ID,KEY_SEQ,SEX,AGE_GROUP,SIDO,RECU_FR_DT,DATA_STD_DT : 관측값들이 넓게 퍼져 있음.(관측값들의 편차가 비교적 큼.)
· 그 외 : 관측값들이 좁은 지역에 밀집되어 있음.(관측값들의 편차가 매우 작음)

2.6.그룹별 관측값 분포의 확인(산점도 이용)

그룹 정보를 이용하여 각 변수별로 그룹별 분포를 확인해보았다.

(1) 산점도를 이용하여 나이그룹과 임내원 일수 사이의 상관관계를 알아보았음.

```
age_group <- diagnosis$AGE_GROUP
recn <- diagnosis$RECN
plot(age_group, recn, main ="나이그룹- 입내원 일수",xlab="나이그룹",ylab="입내원 일수",col="red",pch=19)
```

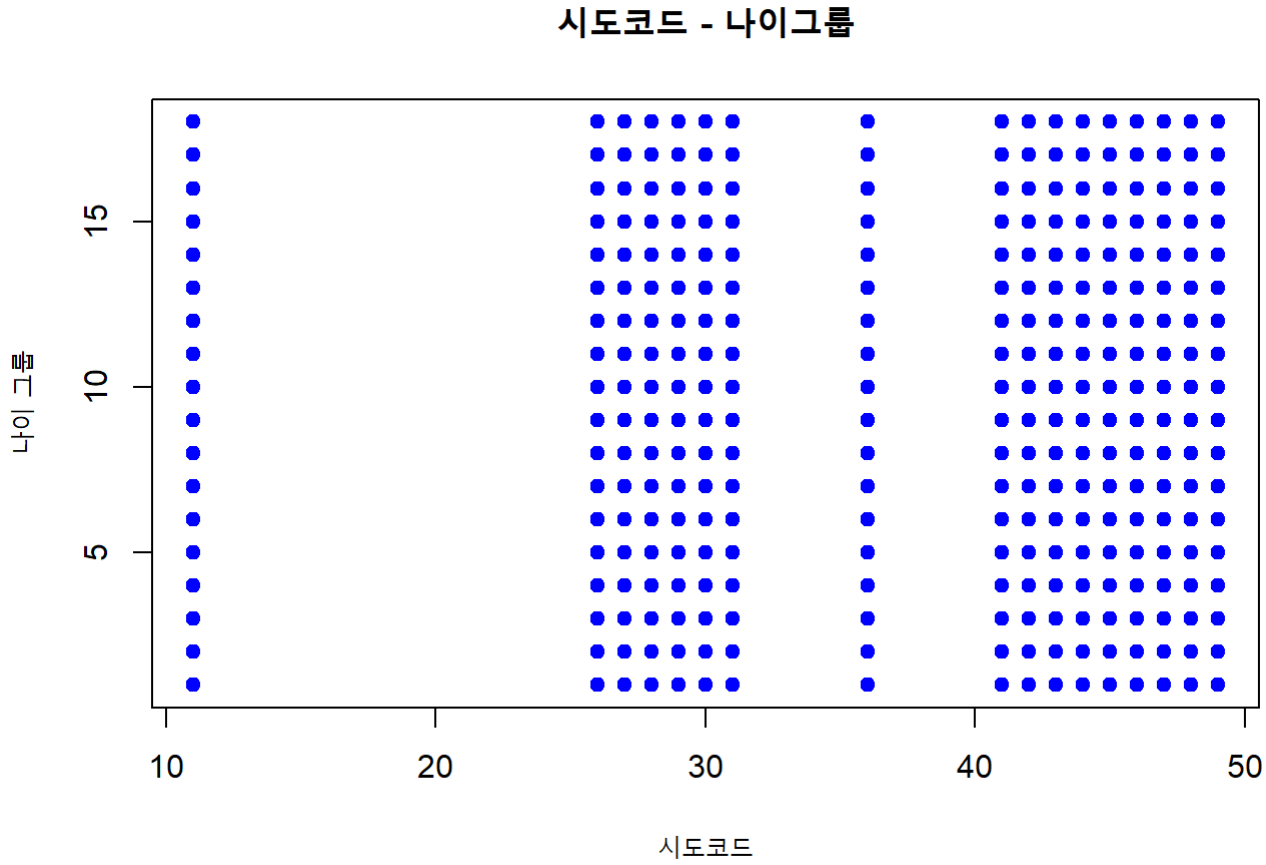


·모든 나이그룹에서 입내원 일수는 대부분 0~200일 사이에 분포함.

·나이가 어릴수록 입내원 일수가 짧다는 것을 볼 수 있음.

(2) 산점도를 통해 시도코드와 나이그룹간의 상관관계를 알아보았음.

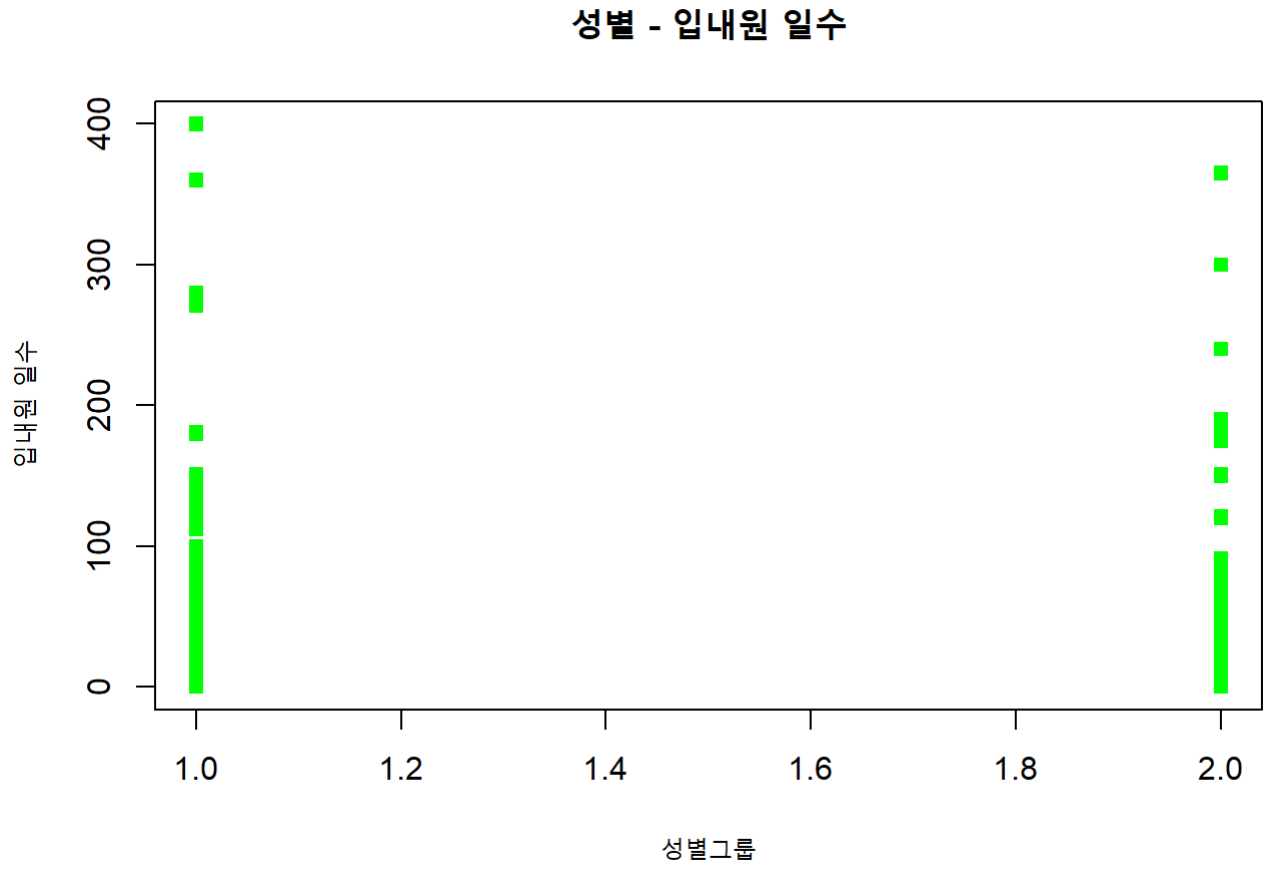
```
age_group <- diagnosis$AGE_GROUP
sido <- diagnosis$SIDO
plot(sido, age_group, main ="시도코드 - 나이그룹",xlab="시도코드",ylab="나이 그룹",col="blue",pch=19)
```



서울,부산,대구,인천 등 우리나라 도시 별로 안과 진료받은 사람들의 연령대는 거의 완전히 고르게 분포되어 있음. 도시별로 더 많이 분포하는 나이그룹을 찾아보려 했지만 균일한 결과값이 나왔음.

(3) 산점도를 통해 성별과 입내원 일수간의 상관관계를 알아보았음.(1 : 남자, 2 : 여자)

```
sex <- diagnosis$SEX
recn <- diagnosis$RECN
plot(sex, recn, main ="성별 - 입내원 일수",xlab="성별그룹",ylab="입내원 일수",col="green",pch=15)
```

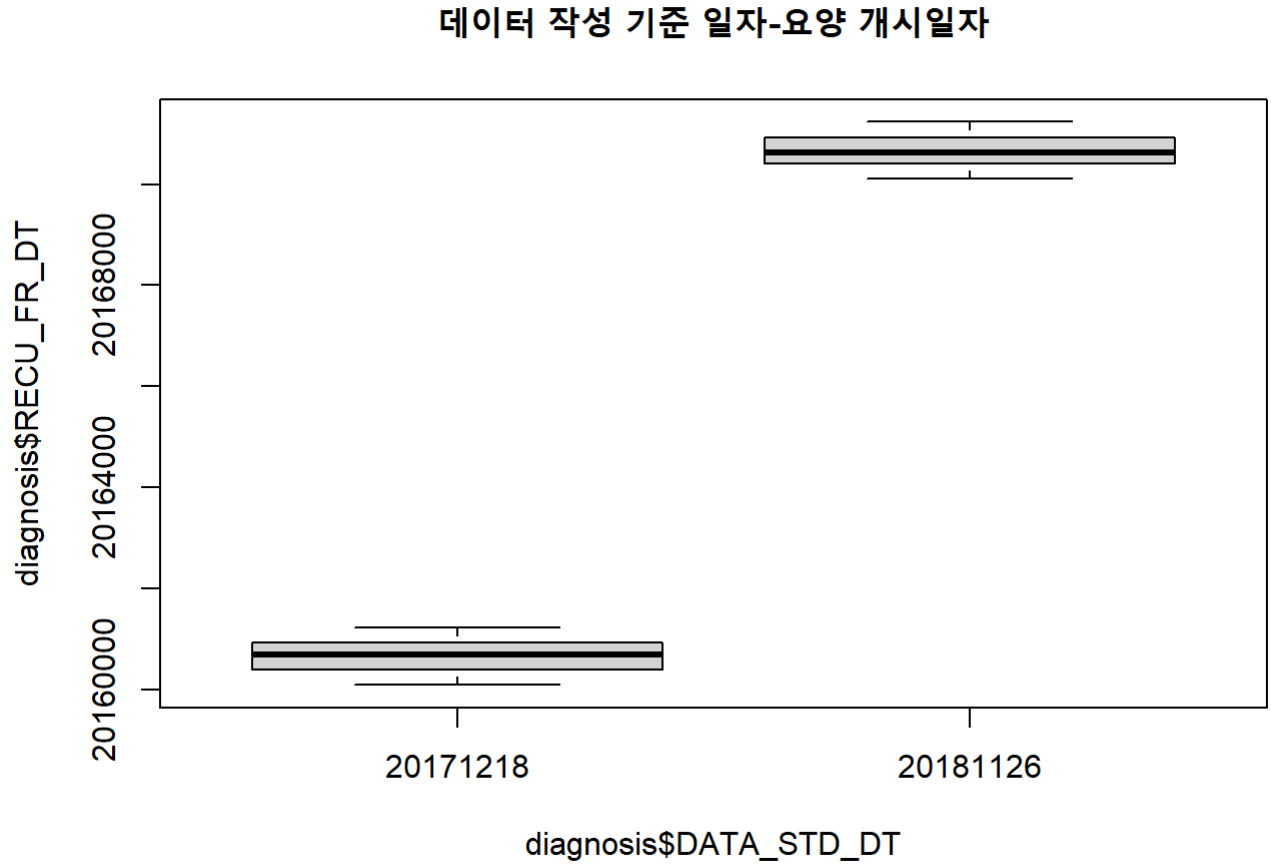


남여 모두 큰 차이없이 0~200일 쪽에 몰려서 분포함.

2.7.그룹별 관측값 분포의 확인(상자그림 이용)

(1) 상자그림을 통해 요양 개시일자과 데이터 작성 기준 일자간의 상관관계를 알아보았음.

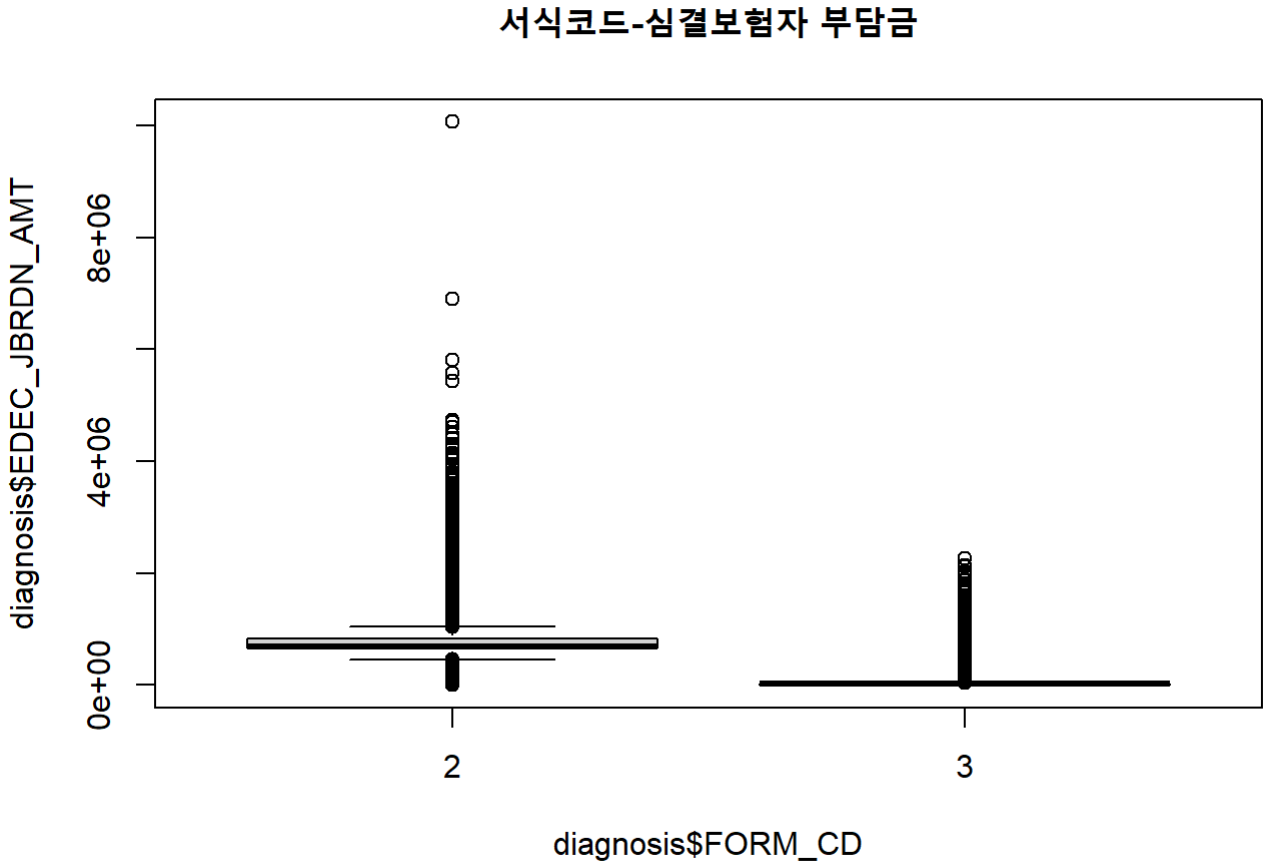
```
boxplot(diagnosis$RECU_FR_DT~diagnosis$DATA_STD_DT, main="데이터 작성 기준 일자-요양 개시일자")
```



· 2017년 12월 18일에 요양을 개시한 경우는 2016년 초에 데이터가 작성되었고, 2018년 11월 26일에 요양을 개시한 경우는 2017년에 데이터가 작성되었다는 것을 알 수 있음.
· 요양을 일찍 받았을수록 데이터 작성이 빨리 되었다는 것을 알 수 있음.

(2) 상자그림을 통해 서식코드와 심결보험자 부담금 간의 상관관계를 알아보았음.

```
boxplot(diagnosis$EDEC_JBRDN_AMT~diagnosis$FORM_CD, main="서식코드-심결보험자 부담금")
```



· FORM_CD(서식코드)가 2일 때, 즉, 의과입원일 때는 심결보험자 부담금이 높고, FORM_CD(서식코드)가 3일 때, 즉, 의과외래일 때는 심결보험자 부담금이 낮게 나타남. 입원이 진료보다 비싼 것은 당연한 결과임.

2.8. 기본함수를 이용하여 최대,최소,평균값, 범위 분석

(1) EDEC_TRAMT(심결요양급여비용 총액)

·최댓값

```
max(diagnosis$EDEC_TRAMT)
```

```
## [1] 12947070
```

·최솟값

```
min(diagnosis$EDEC_TRAMT)
```

```
## [1] 40
```

·평균값

```
mean(diagnosis$EDEC_TRAMT)
```

```
## [1] 38072.34
```

·범위

```
range(diagnosis$EDEC_TRAMT)
```

```
## [1]       40 12947070
```

·EDEC_TRAMT(심결요양급여비용 총액)의 최댓값과 최솟값, 평균값, 범위는 위와 같고 최댓값과 최솟값이 엄청난 차이가 남.

(2) TOT_PRES_DD_CNT(총 처방일수)의 최댓값, 최솟값, 평균값, 범위 분석

·최댓값

```
max(diagnosis$TOT_PRES_DD_CNT)
```

```
## [1] 756
```

·최솟값

```
min(diagnosis$TOT_PRES_DD_CNT)
```

```
## [1] 0
```

·평균값

```
mean(diagnosis$TOT_PRES_DD_CNT)
```

```
## [1] 4.018733
```

·범위

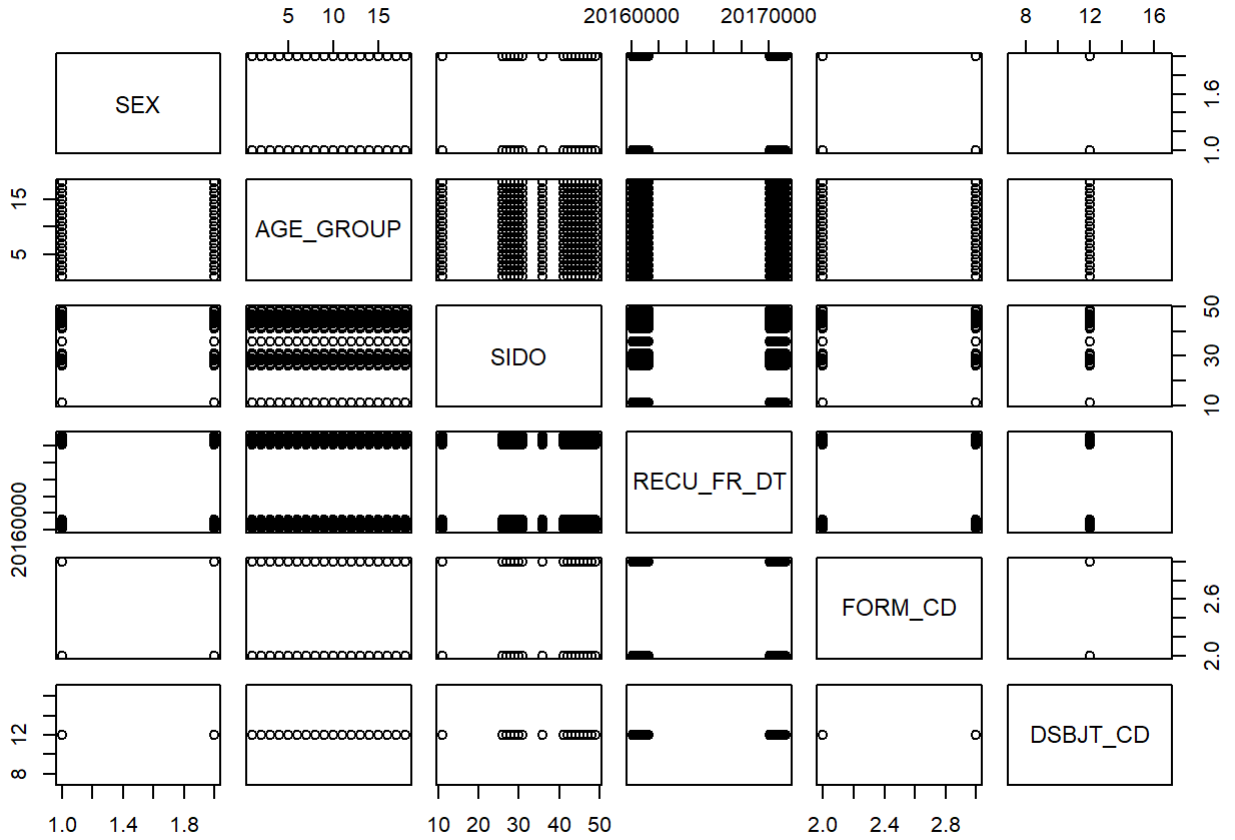
```
range(diagnosis$TOT_PRES_DD_CNT)
```

```
## [1]   0 756
```

·TOT_PRES_DO_CNT(총 처방일수)의 최댓값, 최솟값, 평균값, 범위는 위와 같음.

2.9.

```
pairs(diagnosis[c(4:9)])
```



2.10. 변수 간 상관계수의 확인

DSBJT_CD은 NA로 뜨기 때문에 제외하고, 나머지의 상관계수를 알아보았다.

```
cor(diagnosis[,-c(9:11)])
```

```
##                      STND_Y        IDV_ID       KEY_SEQ           SEX
## STND_Y          1.000000000 -0.0022989650 -0.0276485798 -0.0002623070
## IDV_ID         -0.002298965  1.0000000000  0.0008694387 -0.0031210730
## KEY_SEQ        -0.027648580  0.0008694387  1.0000000000  0.0014090660
## SEX            -0.000262307 -0.0031210730  0.0014090660  1.0000000000
## AGE_GROUP       0.014031730 -0.0029273963 -0.0154058059  0.0529273540
## SIDO           -0.001218899 -0.0015526309 -0.0013905758 -0.0111778170
## RECU_FR_DT      0.997698295 -0.0023104803  0.0090529217 -0.0007419511
## FORM_CD        -0.002544442  0.0004333570  0.0190754037  0.0052438957
## VSCN            0.001565635 -0.0005063294 -0.0042520499 -0.0040701160
## RECN            0.001776267 -0.0007949442 -0.0030745909 -0.0326837285
## EDEC_ADD_RT     0.002250534 -0.0011967067 -0.0374299796 -0.0492599520
## EDEC_TRAMT      0.016288927 -0.0002590782 -0.0183881982 -0.0179082908
## EDEC_SBRDN_AMT  0.022469132  0.0003169109 -0.0215931877 -0.0194862831
## EDEC_JBRDN_AMT  0.013787791 -0.0003703855 -0.0165722108 -0.0151511816
## TOT_PRES_DD_CNT 0.004330069 -0.0004479033  0.0101203336  0.0008347842
## DATA_STD_DT     1.000000000 -0.0022989650 -0.0276485798 -0.0002623070
##                    AGE_GROUP          SIDO    RECU_FR_DT      FORM_CD
## STND_Y          0.014031730 -0.0012188989  0.9976982955 -0.002544442
## IDV_ID         -0.002927396 -0.0015526309 -0.0023104803  0.000433357
## KEY_SEQ        -0.015405806 -0.0013905758  0.0090529217  0.019075404
## SEX             0.052927354 -0.0111778170 -0.0007419511  0.005243896
## AGE_GROUP       1.000000000  0.0132344774  0.0130713482 -0.041011793
## SIDO            0.013234477  1.0000000000 -0.0013629743 -0.004799574
## RECU_FR_DT      0.013071348 -0.0013629743  1.0000000000 -0.002369041
## FORM_CD        -0.041011793 -0.0047995736 -0.0023690408  1.000000000
## VSCN            0.009531086  0.0016030671  0.0015827979 -0.289719246
## RECN            0.032967488 -0.0151798039  0.0018521286 -0.073756993
## EDEC_ADD_RT     0.073194054 -0.0536299476  0.0025264048 -0.060343132
## EDEC_TRAMT      0.072781623 -0.0031131631  0.0164384211 -0.715434090
## EDEC_SBRDN_AMT  0.070587701 -0.0167832456  0.0227925138 -0.591594405
## EDEC_JBRDN_AMT  0.069475981  0.0009913924  0.0138759367 -0.725193247
## TOT_PRES_DD_CNT 0.088093318  0.0177826892  0.0045734478  0.023595540
## DATA_STD_DT     0.014031730 -0.0012188989  0.9976982955 -0.002544442
##                         VSCN          RECN  EDEC_ADD_RT    EDEC_TRAMT
## STND_Y          0.0015656348  0.0017762666  0.002250534  0.0162889265
## IDV_ID         -0.0005063294 -0.0007949442 -0.001196707 -0.0002590782
## KEY_SEQ        -0.0042520499 -0.0030745909 -0.037429980 -0.0183881982
## SEX            -0.0040701160 -0.0326837285 -0.049259952 -0.0179082908
## AGE_GROUP       0.0095310863  0.0329674881  0.073194054  0.0727816226
## SIDO            0.0016030671 -0.0151798039 -0.053629948 -0.0031131631
## RECU_FR_DT      0.0015827979  0.0018521286  0.002526405  0.0164384211
## FORM_CD        -0.2897192457 -0.0737569930 -0.060343132 -0.7154340898
## VSCN            1.0000000000  0.1223914809  0.049155250  0.4000019459
## RECN            0.1223914809  1.0000000000  0.103003420  0.1104000851
## EDEC_ADD_RT     0.0491552496  0.1030034196  1.000000000  0.1868426011
## EDEC_TRAMT      0.4000019459  0.1104000851  0.186842601  1.0000000000
## EDEC_SBRDN_AMT  0.3625361642  0.0857904131  0.358520482  0.8974060763
## EDEC_JBRDN_AMT  0.3957984292  0.1043047601  0.126134711  0.9916844098
## TOT_PRES_DD_CNT -0.0058083284  0.0448105830  0.120218474  0.0128206196
## DATA_STD_DT     0.0015656348  0.0017762666  0.002250534  0.0162889265
##                 EDEC_SBRDN_AMT EDEC_JBRDN_AMT TOT_PRES_DD_CNT  DATA_STD_DT
## STND_Y           0.0224691320   0.0137877905    0.0043300694  1.000000000
## IDV_ID           0.0003169109  -0.0003703855   -0.0004479033 -0.002298965
## KEY_SEQ         -0.0215931877  -0.0165722108    0.0101203336 -0.027648580
```

```
## SEX              -0.0194862831  -0.0151511816    0.0008347842 -0.000262307
## AGE_GROUP         0.0705877013   0.0694759806    0.0880933184  0.014031730
## SIDO             -0.0167832456   0.0009913924    0.0177826892 -0.001218899
## RECU_FR_DT        0.0227925138   0.0138759367    0.0045734478  0.997698295
## FORM_CD          -0.5915944047  -0.7251932473    0.0235955404 -0.002544442
## VSCN              0.3625361642   0.3957984292   -0.0058083284  0.001565635
## RECN              0.0857904131   0.1043047601    0.0448105830  0.001776267
## EDEC_ADD_RT       0.3585204815   0.1261347112    0.1202184744  0.002250534
## EDEC_TRAMT        0.8974060763   0.9916844098    0.0128206196  0.016288927
## EDEC_SBRDN_AMT    1.0000000000   0.8353191924    0.0456476162  0.022469132
## EDEC_JBRDN_AMT    0.8353191924   1.0000000000    0.0022248978  0.013787791
## TOT_PRES_DD_CNT   0.0456476162   0.0022248978    1.0000000000  0.004330069
## DATA_STD_DT       0.0224691320   0.0137877905    0.0043300694  1.000000000
```

· RECU_FR_DT를 기준으로 보았을 때 상관계수가 가장 높은 것은 DATA_STD_DT으로, 0.9976952955임. (2.7.(1)의 내용 참고)

· FORM_CD를 기준으로 보았을 때 음의 상관성이 높은 것은 EDEC_JBRDN_AMT으로, -0.725193247임. (2.7.(2)의 내용 참고)

· EDEC_TRAMENT를 기준으로 보았을 때 상관계수가 가장 높은 것은 EDEC_JBRDN_AMT으로, 0.9916844098이고, 음의 상관성이 높은 것은 FORM_CD로, -0.7154340898임.

· EDEC_SBRDN_AMT을 기준으로 보았을 때 상관계수가 가장 높은 것은 EDEC_TRAMT으로, 0.8974060763임.

· EDEC_JBRDN_AMT를 기준으로 보았을 때 음의 상관성이 가장 높은 것은 FORM_CD로, -0.7251932473임.

=> EDEC_TRAMENT(심결요양 급여비용 총액) = EDEC_SBRDN_AMT(심결본인 부담금) + EDEC_JBRDN_AMT(심결 보험자 부담금)이므로 세 변수간의 양의 상관성은 높을 수 밖에 없음.

· 2.6.(1)에서 양의 상관성이 높다고 보였던 AGE_GROUP(나이그룹)과 RECN(입원일수)는 상관계수가 0.032967488으로, 실제로는 높지 않다는 것을 알 수 있음.