2021 YDMS 5기 2차 모집과제

정보통계학과 2019251034 김하은

<1. 본인이 생각하는 EDA란 어떠한 것인지 서술하시오.>

우선 EDA가 무엇인지에 대해 자세히 알아보기 앞서, 용어가 대강 어떤 의미를 내포하고 있는지에 대해 생각해보았다. EDA 란 Exploratory Data Analysis 의 약어로 '탐색적 자료 분석'을 뜻한다. 탐색이란, 드러나지 않은 (사물이나) 현상 따위를 찾아내거나 밝히기 위하여 살피어 찾는다는 의미이다. 그러므로 탐색적 자료 분석이란 자료의 숨겨진 정보 및 데이터 분포에 대한 패턴을 찾아내기 위해, 자료를 살피어 분석하는 방식이라고 생각하였다.

EDA를 창안한 통계학자 존 튜키에 따르면, EDA는 가설 검정이 주목적인 기존 통계학과는 달리 데이터 속 정보를 파악하는 것에 주목적을 둔 분석 방식으로, 주어진 자료를 가지고 정보를 찾아내는 방식이다.

다시 말해, EDA는 가설 검정에 쏠려 자료의 본질을 찾기 힘들었던 기존 통계학의 문제점을 보완하는 방법이기에, 주어진 자료를 탐색하여 숨겨진 자료 본연의 정보를 찾는데에 집중한다. 이때 EDA는 다양한 각도로 데이터를 탐색함으로써 데이터에 대한 더 깊은 이해를 도울 수 있다. 또한 숨겨진 자료 속 문제점을 발견하게 되는 경우, 문제점을 수정하여 재탐색하는 과정이 가능하다는 점은 EDA의 큰 장점이다.

한편 EDA를 통해 자료 본연의 의미를 정확하게 파악하기 위해서는 변수에 대한 이해가 반드시 되어있어야 한다. 따라서 의미있는 정보를 얻기 위해선 변수끼리의 조합을 다양하게 고려해보는 것은 물론 변수끼리의 관계가 잘 맺어진 것인지도 반드시 확인해보아야 한다. 그래야만 EDA를 통해 자료 속 의미있는 패턴과 정보를 찾아낼 수 있기 때문이다. 만약 변수 유형을 무시한 채 진행하게 되면 전혀 다른 내용의 정보가 나올 수 있기 때문에 변수 유형에 주의해야한다.

마지막으로 EDA의 과정은 다음과 같다. 우선 자료 속 변수에 대한 이해를 바탕으로 시작되어야 한다. 자료의 관측값들을 확인해보는 것도 좋고, 어떠한 형태인지, 변수 유형은 어느 쪽에 속하는지, 변수가 의미하는 것은 무엇인지에 대해 파악하는 것이 여기에 해당된다. 이후 자료를 살피면서 이상치나 결측값과 같은 문제점이 없는지 따져가며 진행해야 하고, 문제점이 있는 경우 자료를 수정한 뒤 탐색하여도 좋다. 이후 자료를 시각화하여 각 변수값들이 어떠한 패턴을 따르는지, 각 변수 간의 관계는 어떻게 되는지에 대해 파악하며 탐색을 진행한다. 이 과정을 통해 데이터 속 숨겨진 패턴이나 정보를 알아낼 수 있으며, 그에 대한 해석하는 과정까지 EDA에 속한다고 할 수 있다.

※ <2번>, <3번> 동시에 진행했습니다!

분석하기에 앞서, 변수에 대한 요약을 해보면 다음과 같다.

순서	변수명	변수 설명	
1	▶male	범주형	0 = Female 1 = Male
2	▶age	연속형	Age at exam time
3	▶education	범주형	1 = Some High School 2 = High School or GED 3 = Some College or Vocational School 4 = college
4	▶currentSmoker	범주형	0 = nonsmoker 1 = smoker
5	▶cigsPerDay	연속형	number of cigarettes smoked per day (estimated average)
6	▶BPMeds	범주형	0 = Not on Blood Pressure medications 1 = Is on Blood Pressure medications
7	▶prevalentStroke	범주형	뇌졸중여부 0 = No 1 = Yes
8	▶prevalentHyp	범주형	우울여부 0 = No 1 = Yes
9	▶diabetes	범주형	당뇨병 여부 0 = No 1 = Yes
10	▶totChol	연속형	콜레스테롤 수치 [단위: mg/dL]
11	▶sysBP	연속형	수축기 혈압 [단위: mmHg]
12	▶diaBP	연속형	이완기 혈압 [단위: mmHg]
13	▶BMI	연속형	Body Mass Index calculated as: Weight (kg) / Height(meter-squared)
14	▶heartRate	연속형	Beats/Min (Ventricular)
15	▶glucose	연속형	클루코즈 [단위: mg/dL]
16	▶TenYearCHD (타겟 변수)	범주형	십년후 관상동맥질환발병 여부 0 = No 1 = Yes

summary 함수를 이용하여 변수 별 형태 및 오류를 알아보았고, 이를 정리하면 다음과 같다.

```
> summary(framingham)
 male        age          education     currentSmoker     cigsPerDay        BPMeds       prevalentStroke
 0:2416   Min.   :32.00   Min.   :1.000   Min.   :0.0000   Min.   :-1.000   Min.   :0.00000   Min.   :0.000000
 1:1815   1st Qu.:42.00   1st Qu.:1.000   1st Qu.:0.0000   1st Qu.: 0.000   1st Qu.:0.00000   1st Qu.:0.000000
 F:   4   Median :49.00   Median :2.000   Median :0.0000   Median : 0.000   Median :0.00000   Median :0.000000
 M:   5   Mean   :49.58   Mean   :1.979   Mean   :0.5349   Mean   : 9.001   Mean   :0.02962   Mean   :0.005896
          3rd Qu.:56.00   3rd Qu.:3.000   3rd Qu.:1.0000   3rd Qu.:20.000   3rd Qu.:0.00000   3rd Qu.:0.000000
          Max.   :70.00   Max.   :4.000   Max.   :2.0000   Max.   :70.000   Max.   :1.00000   Max.   :1.000000
                          NA's   :105                      NA's   :29       NA's   :53
 prevalentHyp    diabetes     totChol          sysBP           diaBP             BMI          heartRate
 Min.   :0.0000   0 :4097   Min.   :107.0   Min.   : 83.5   Min.   : 48.00   Min.   :15.54   Min.   : 44.00
 1st Qu.:0.0000   1 : 109   1st Qu.:206.0   1st Qu.:117.0   1st Qu.: 75.00   1st Qu.:23.07   1st Qu.: 68.00
 Median :0.0000   No:  34   Median :234.0   Median :128.0   Median : 82.00   Median :25.40   Median : 75.00
 Mean   :0.3106             Mean   :236.7   Mean   :132.4   Mean   : 82.89   Mean   :25.80   Mean   : 75.88
 3rd Qu.:1.0000             3rd Qu.:263.0   3rd Qu.:144.0   3rd Qu.: 90.00   3rd Qu.:28.04   3rd Qu.: 83.00
 Max.   :1.0000             Max.   :696.0   Max.   :295.0   Max.   :142.50   Max.   :56.80   Max.   :143.00
                            NA's   :50                      NA's   :30       NA's   :19      NA's   :1
    glucose        TenYearCHD
 Min.   :  40.00   .:   3
 1st Qu.:  71.00   0:3593
 Median :  78.00   1: 644
 Mean   :  88.99
 3rd Qu.:  87.00
 Max.   :5500.00
 NA's   :383
```

순서	변수명	이상값 존재 여부
1	▶male	▶0,1 이외의 F,M 존재
2	▶age	
3	▶education	▶ 결측치 105개
4	▶currentSmoker	▶0,1 이외의 2존재
5	▶cigsPerDay	▶결측치 29개, 음수값 존재
6	▶BPMeds	▶결측치 53개
7	▶prevalentStroke	
8	▶prevalentHyp	
9	▶diabetes	▶0,1 이외의 NO 존재
10	▶totChol	▶결측치 50개
11	▶sysBP	
12	▶diaBP	▶결측치 30개
13	▶BMI	▶결측치 19개
14	▶heartRate	▶결측치 1개
15	▶glucose	▶결측치 383개
16	▶TenYearCHD	▶0,1 이외의 . 존재, 이상점 존재

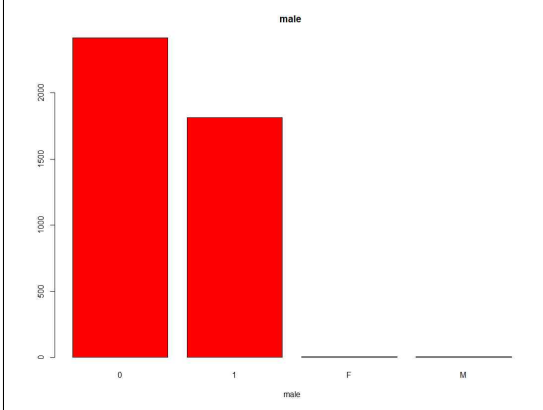
타겟 변수를 보아, 15가지 변수의 흐름과 10년 뒤 관상동맥 질병의 발현 여부를 비교하여 각 변수들이 어떠한 관계를 맺고 있으며, 이들이 'TenYearCHD'에는 어떠한 영향을 미치는지를 파악하는 것이 이번 과제의 목표이다.

#male

male의 경우 성별을 나타내는 변수로 0 또는 1 의 값을 가지고 있어야 한다. 하지만 주어진 자료에는 F나 M과 같은 값들이 존재했기 때문에 이를 수정해주었다. 삭제하는 방향보단, 통상 쓰이는 관념에 맞춰 수정하였다. F의 경우 Female 이라 보고 0으로 수정하였으며, M의 경우 Male 이라 보고 1로 수정하였다. 수정 후 분포는 아래의 그래프와 같다.

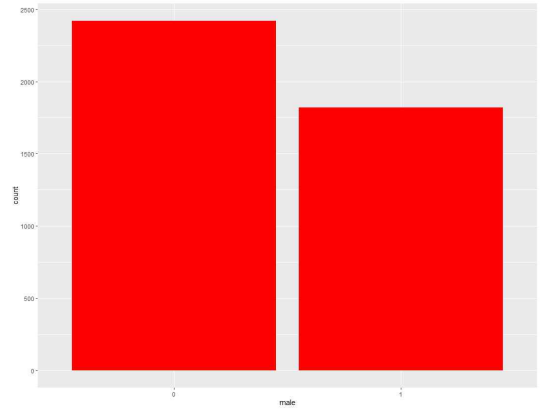
한편 수정 후 테이블을 이용하여 각 영역의 도수를 알아보았는데, 여성의 경우 2420명, 남성의 경우 1820명임을 알 수 있다. 여성이 남성보다 600명 정도 많다.

수정 전




```
count<-table(fram$male)
barplot(count,xlab="male",main = "male ",col="red")
```

수정 후

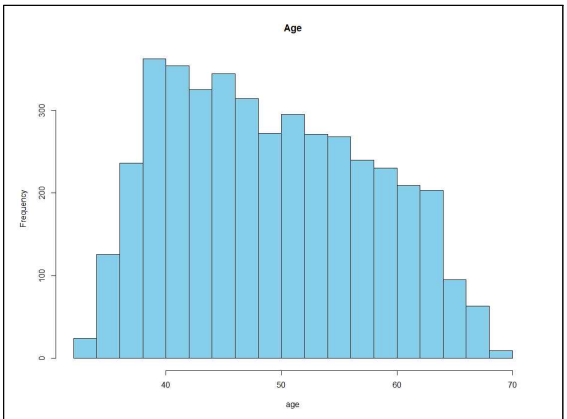



```
replace(fram$male,fram$male=="F",0)->fram$male
replace(fram$male,fram$male=="M",1)->fram$male
ggplot(fram, aes(male))+geom_bar(fill='red')
```

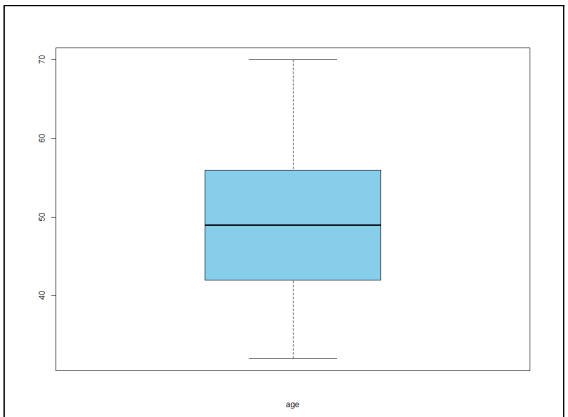
```
> summary(fram$male)
   0    1    F    M
2420 1820    0    0
```

#age

age의 경우 나이를 나타내는 변수이다. 수정해야할 부분이 없었기에 바로 그래프를 그렸고, 이때 연속형 변수임을 고려하여 히스토그램과 상자그림을 그렸다. 그래프를 통해 이 자료는 전연령을 대상으로 하는 것이 아닌 30대에서 70대를 대상으로 한 자료임을 알 수 있었다.




```
age<-fram$age
hist(age,col="sky blue",main = "Age")
```




```
boxplot(fram$age,horizontal=F, col="sky blue",xlab="age")
```

#education

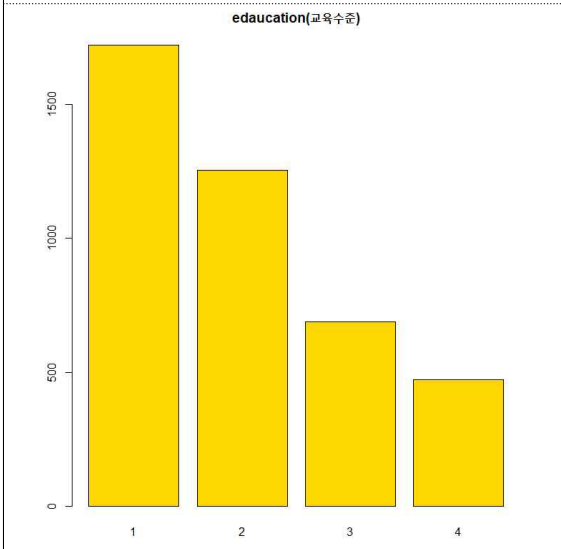
education 은 교육의 수준을 범주화 시켜 나눈 변수로 105개의 결측값을 가지고 있다. 우선 다른 변수와의 관계를 살펴보았는데 큰 연관성을 띠고 있는 변수가 없다고 판단하였다. 하지만 결측값이 105개 정도로 삭제하기엔 많다고 생각하여 그룹5 로 다시 묶어 그룹화하였다. 즉 education =5 인 그룹은 교육수준이 미정인 그룹인 것이다.

이후 수정한 테이블을 바탕으로 막대그래프를 그려보았다. 수정 전 막대그래프와 비교하면 그룹 5가 추가된 것 이외엔 큰 변동사항이 없다. 또한 그룹별 도수를 살펴보면 그룹 5 전체가 다른 하나의 그룹에 속해질지라도 막대그래프의 변화 모양에는 큰 변화가 없음을 알 수 있다. 그만큼 그룹 5의 도수는 상대적으로 작은 편이다. 따라서 그룹 5로 인한 데이터 오류는 크지않을 것이라 판단하여 그대로 진행하였다.

```
> fram[is.na(fram$education),]
     male age education currentSmoker cigsPerDay BPMeds prevalentStroke prevalentHyp diabetes totChol sysBP diaBP   BMI heartRate glucose TenYear
34      1  61        NA             2          5      0               0            0       No      175 134.0  82.5 18.59        72      75
37      1  56        NA             0          0      0               0            0        0      257 153.5 102.0 28.09        72      75
73      0  37        NA             0          0      0               0            0        0      200 119.0    NA 33.29        67      87
185     1  67        NA             0          0      0               0            0        0      257 125.0  67.5 25.95        65      69
214     0  34        NA             0          0      0               0            0        0      163 107.0  71.0 23.88        73      80
294     0  45        NA             1         30      0               0            0        0      203 131.0  85.0 23.47        94      70
306     1  36        NA             1         20      0               0            1        0      304 118.0  90.0 32.63        71      80
307     0  52        NA             0          0      0               0            0        0      268 109.0  70.0 23.74        75      78
320     0  37        NA             1         20      0               0            0        0      223 115.0  72.0 22.71        76      63
401     1  56        NA             1         25      0               0            0        0      255 138.0  80.0 23.44        67      79
413     0  46        NA             1         20      0               0            0        0      212 122.5  75.5 23.51        67     103
430     1  65        NA             0          0      0               0            0        0       NA 152.5  97.5 28.35        65      73
473     0  38        NA             1          1      0               0            0        0      300 122.0  84.0 27.26        96      68
500     0  60        NA             0          0      0               0            0        0      215 113.0  71.0 26.69        77      NA
504     1  38        NA             1         25      0               0            1        0      210 145.5  87.0 24.67        72      89
623     0  64        NA             0          0      0               0            0        0      293 116.0  80.0 26.81        80      87
695     0  38        NA             1         20      0               0            0        0      199 117.0  78.5 18.18        90      73
720     0  38        NA             1          5      0               0            0        0      190 121.0  79.0 25.59        90      84
738     1  48        NA             0          0      0               0            0        0      222 113.0  71.5 30.50        78      80
```

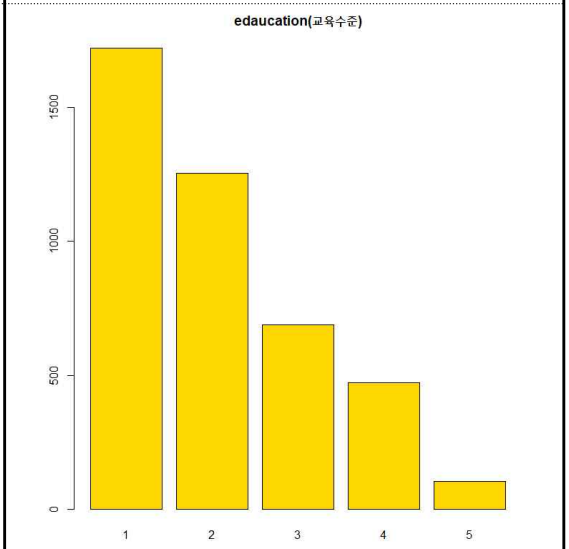
*결과값이 너무 많아서 뒷부분은 생략함 / education 관련 변수를 찾기 힘듦.

수정 전




```
barplot(table(fram$education),col='gold',main ='edaucation(교육수준)')
```

수정 후




```
fram[is.na(fram$education),]
fram$education[is.na(fram$education)]<-5 #NA를 5로 대체
barplot(table(fram$education),col='gold',main ='edaucation(교육수준)')
```

#current Smoker

current smoker는 현재 흡연 여부를 묻는 것으로 흡연자와 비흡연자를 나누는 변수이다. 따라서 current smoker와 cigs per day가 서로 연관되어있을 것이라 생각하였다. cigs per day의 경우 평균적으로 하루에 피는 담배의 수를 나타내는 변수이기 때문에 비흡연자의 경우 무조건 0일 것이라고 판단하였다. 이러한 가정이 맞는지 확인하기 위해 current smoker=0 인 행 추출하였고, current smoker=0 인 변수는 모두 cigs per day=0 임을 확인하였다. 따라서 'current smoker=0 이면 cigs per day=0 이다.' 라는 논리에 확신을 가지고 명제의 대우를 이용하여 'cig per day가 0이 아닌 변수는 current smoker도 0이 아닌 1이다' 라는 논리을 세워 진행하였다. 위의 가정을 가지고 current smoker=2인 변수의 cigs per day를 확인하였고, cigs per day가 모두 0이 아님을 확인하였다. 따라서 위의 논리를 바탕으로 current smoker=2는 1로 수정하였다.

수정한 테이블을 바탕으로 current smoker에 관한 막대그래프를 그렸으며 이는 아래와 같다. 자료 속 흡연자와 비흡연자의 비율은 거의 비슷함을 알 수 있다.

current smoker=0

currentSmoker	cigsPerDay
0	0
0	0
0	0
0	0
0	0
0	0
0	0
0	0
0	0
0	0
0	0
0	0
0	0
0	0

*결과값이 너무 많아서 뒷부분은 생략함.

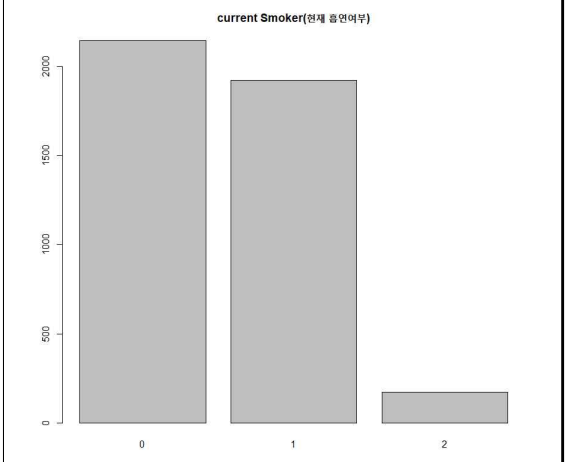
fram[fram$currentSmoker==0,]

current smoker=2

currentSmoker	cigsPerDay
2	30
2	5
2	40
2	3
2	-1
2	20
2	20
2	20
2	15
2	30
2	20
2	20
2	20

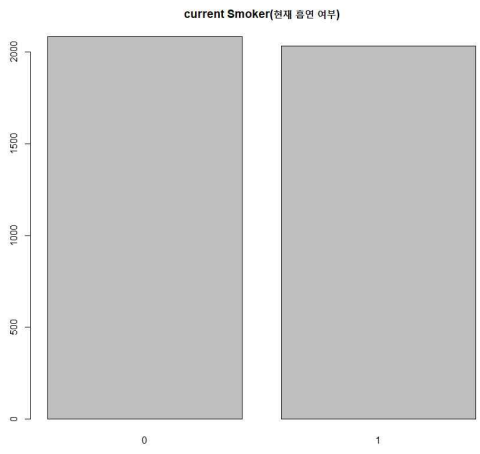
fram[fram$currentSmoker==2,]

수정 전



barplot(table(fram$currentSmoker),col='gray',main='cigs per day(현재 흡연여부)')

수정 후



replace(fram$currentSmoker,fram$currentSmoker==2,1)->fram$currentSmoker **#2값을 1로 수정**

barplot(table(fram$currentSmoker),col='gray',main='cigs per day(현재 흡연 여부)')

#cigs per day

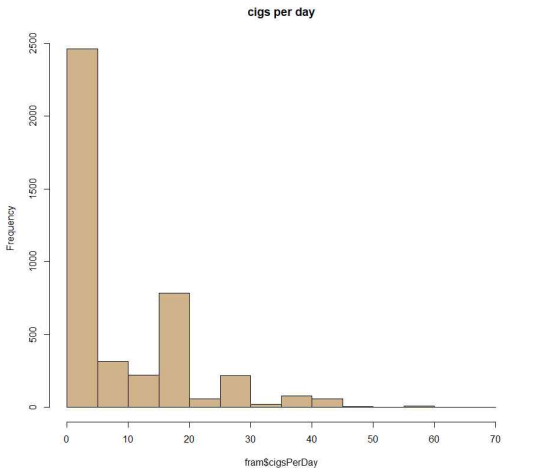
앞서 언급했듯이 cigs per day는 하루 평균 흡연량(담배수)를 의미하는 변수이다. 따라서 이 변수는 절대로 음수가 나올 수 없다. 하지만 테이블에서 이 변수에 대해 오름차순 해본 결과 음수가 있음을 알게 되었고 이를 절댓값 함수를 이용하여 수정해주었다. 수정이 잘 되었는지 확인하기 위해 테이블에서 변수 오름차순을 다시 해보았고 0부터 오름차순 됨을 확인하였다.

수정 전	수정 후
★cigsperday 변수 오름차순으로 비교한 것.	
cigsPerDay -1 -1 -1 -1 -1 -1 -1 -1 -1 -1	cigsPerDay 0 0 0 0 0 0 0 0 0 0
View(fram)	abs(fram$cigsPerDay)->fram$cigsPerDay View(fram)

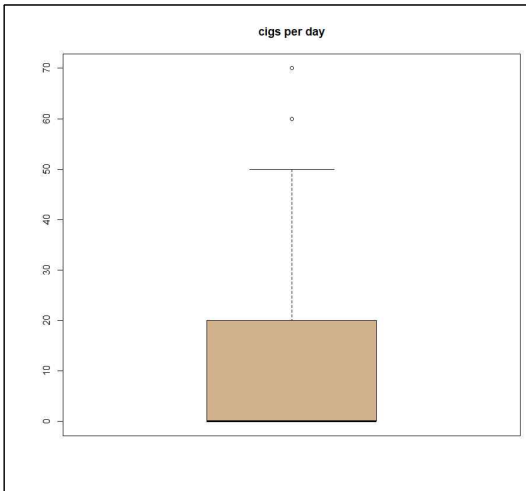
한편 이 변수에는 결측값도 존재한다. 결측값이 존재하는 행들을 모두 검색해본 결과 모두 currentsmoker=1 인 경우였다. 앞서 언급한 것처럼 cigs per day은 currentsmoker와 연관되어있다. currentsmoker=1 이라는 점은 모두 흡연자인 경우이므로 cigs per day에는 흡연자들의 하루 평균 개수를 적용하여 값을 대체하였다.

수정 전	수정 후
male age education currentSmoker cigsPerDay 132 1 43 4 1 NA 140 1 49 4 1 NA 1047 0 49 1 1 NA 1293 1 42 3 1 NA 1348 0 58 4 1 NA 1452 1 54 1 1 NA 1498 1 55 1 1 NA 1611 0 61 1 1 NA 1626 0 49 2 1 NA	male age education currentSmoker cigsPerDay 132 1 43 4 1 18 140 1 49 4 1 18 211 1 45 4 1 18 544 1 47 2 1 18 583 0 51 2 1 18 652 0 56 1 1 18 690 1 64 1 1 18 1047 0 49 1 1 18 1293 1 42 3 1 18 1348 0 58 4 1 18 1452 1 54 1 1 18 1498 1 55 1 1 18 1601 1 66 4 1 18 1611 0 61 1 1 18 1626 0 49 2 1 18
fram[is.na(fram$cigsPerDay),]	fram[fram$currentSmoker==1,]->smk smk.mean<-mean(smk$cigsPerDay,na.rm=TRUE) round(smk.mean)
*결과값이 너무 많아서 뒷부분은 생략함. / 행 번호를 기준으로 비교하면 결측값이 올바르게 바뀐 것을 확인 할 수 있음.	

수정한 내용을 바탕으로 cigs per day의 분포를 파악하기 위해 히스토그램과 상자그림을 그렸다. 해당 그래프의 경우 흡연자와 비흡연자가 모두 포함된 상태에서 그려진 것이라 0에 가장 많은 분포를 하고 있다. 흡연자와 비흡연자 모두 포함된 상태에서의 평균 cigs per day는 9.063897이다.

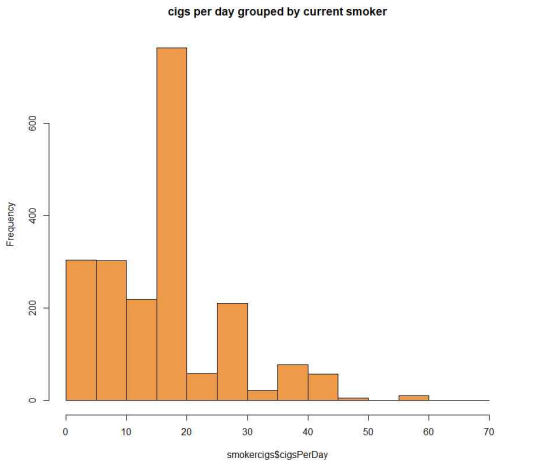


hist(fram$cigsPerDay,col='tan',main="cigs per day")

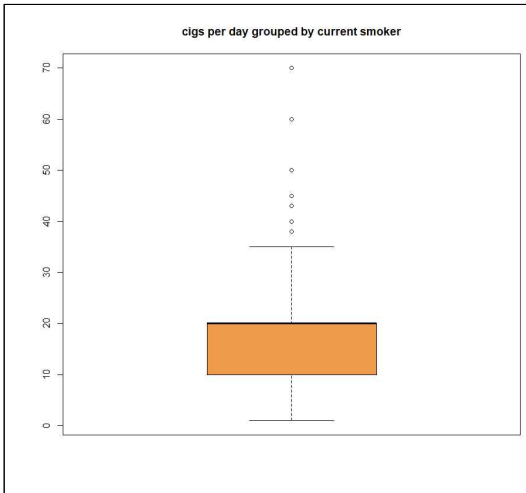


boxplot(fram$cigsPerDay,col='tan',main="cigs per day")

흡연자와 비흡연자를 모두 포함시킨 cigs per day의 분포도 좋지만 흡연자만의 cigs per day 분포를 파악하는 것도 분석에 유의한 결과를 내올 수 있을 것 같아 흡연자만의 cigs per day를 따로 분석해보았다. 분석해본 결과 평균은 18.36으로 두 배 가량 증가하였고, 그래프의 분포양상에도 큰 변화가 생김을 알 수 있다.



hist(smokercigs$cigsPerDay,col='tan2',main="cigs per day grouped by current smoker")



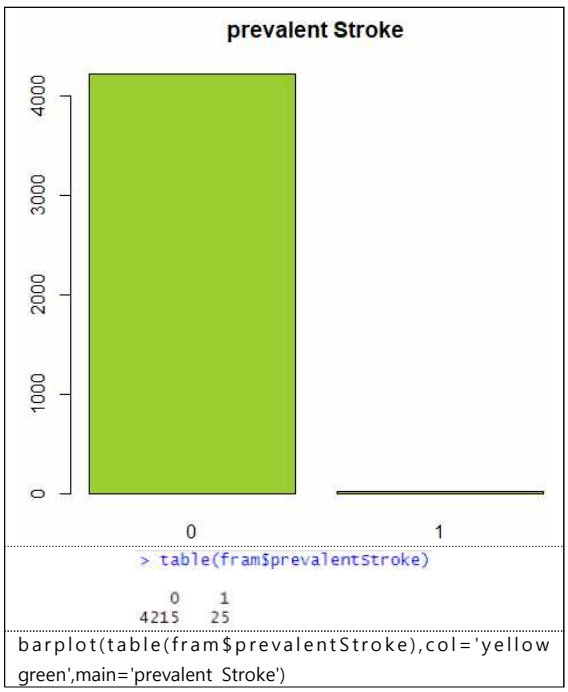
boxplot(smokercigs$cigsPerDay,col='tan2',main="cigs per day grouped by current smoker")

#BPMeds

BPMeds는 혈압량 복용 여부에 대한 변수로 결측값을 53개 가지고 있다. 이 변수의 경우 후술할 연속형 변수의 수정을 바탕으로 분석하였기 때문에 연속형 변수 분석 이후 다시 서술하였다. (p14)

#prevalent Stroke

prevalent Stroke는 뇌졸중 여부에 대한 변수이다. 수정해야할 오류사항이 없었기 때문에 범주형 변수에 맞게 막대그래프를 그려 분포를 살펴보았다. 압도적으로 뇌졸중을 가진 환자가 적었으며 4240명 중 25명만이 이 질환을 겪은 것으로 나타났다. 이 외에 특별한 사항을 찾을 수 없었다.

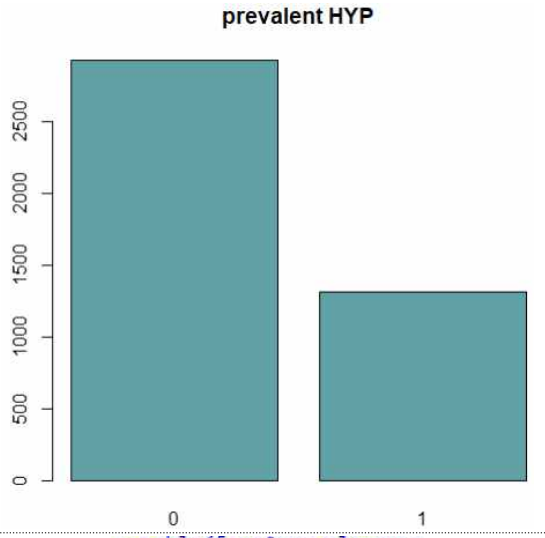


#prevalent HYP

prevalent HYP는 우울 여부에 관한 변수이다. 마찬가지로 수정해야할 오류사항이 없었기 때문에 범주형 변수에 맞게 막대그래프를 그려 분포를 살펴보았다. 우울 증상이 없는 사람이 더 많았지만 우울증상을 나타내는 사람도 4240명 중 1317명으로 꽤 높은 비율을 차지함을 알 수 있다.

혈압량 복용여부를 나타내는 BPMeds에서 후술할 내용이지만 prevalent HYP와 BPMeds는 연관성이 있다. 다시말해, 혈압약을 복용하는 환자의 경우 우울증상이 동반된다는 것이다. 실제로 고혈압 약이 우울증을 유발할 수 있다는 연구결과가 있다.[1)]

1) https://www.pharmnews.com/news/articleView.html?idxno=84031



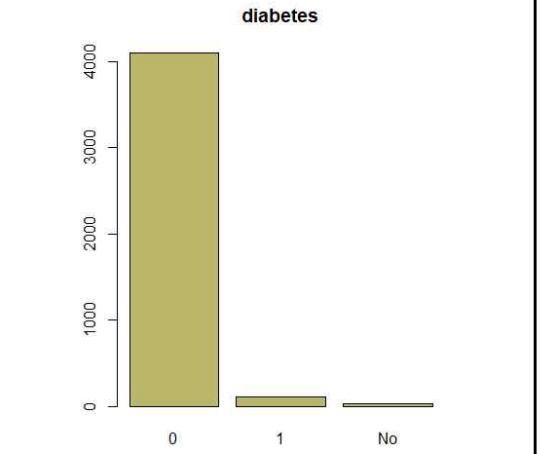

```
> table(fram$prevalentHyp
   0    1
2923 1317
```

```
barplot(table(fram$prevalentHyp),col='cadet blue',main='prevalent HYP')
```

#diabetes

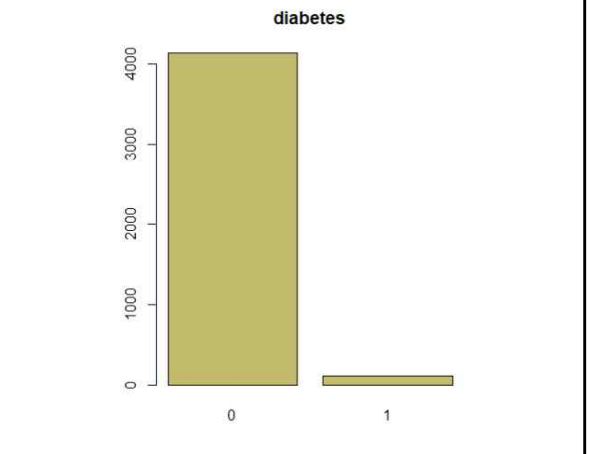
diabetes의 경우 당뇨병 질환 여부를 나타내는 변수이다. 이 변수는 0,1 이외의 NO를 가지고 있기 때문에 NO에 대한 값을 수정하였다. NO를 1로 수정하였고 이를 바탕으로 그래프를 다시 그려보았다. 그래프 분포에는 큰 영향이 없지만 1에 관한 도수가 늘어났다.

수정 전




```
barplot(table(fram$diabetes),col='dark khaki',main='diabetes')
```

수정 후




```
fram$diabetes<-gsub("No","0",fram$diabetes) #No를 0으로 수정
barplot(table(fram$diabetes),col='dark khaki',main='diabetes')
```

#연속형 변수들 사이의 상관관계

연속형 변수들을 다루기에 앞서, 연속형 변수들만 따로 빼내어 테이블을 만들었고, 연속형 변수 간 상관관계를 따져보았다. 이때 결측값이 하나라도 있는 경우 계산이 되지 않기 때문에 결측값이 있는 행은 제외하고 상관관계를 따져보았다.

<연속형 변수만 모아놓은 테이블>

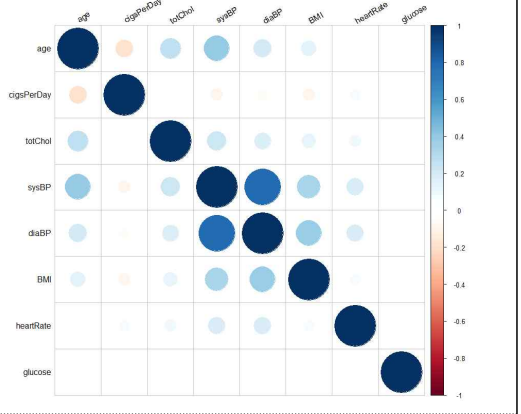
	age	cigsPerDay	totChol	sysBP	diaBP	BMI	heartRate	glucose
1	39	0	195	106.0	70.0	26.97	80	77
2	46	0	250	121.0	81.0	28.73	95	76
3	48	20	245	127.5	80.0	25.34	75	70
4	61	30	225	150.0	95.0	28.58	65	103
5	46	23	285	130.0	84.0	23.10	85	85
6	43	0	228	180.0	110.0	30.30	77	99
7	63	0	205	138.0	71.0	33.11	60	85
8	45	20	313	100.0	71.0	21.68	79	78
9	52	0	260	141.5	89.0	26.36	76	79
10	43	30	225	162.0	107.0	23.61	93	88
11	50	0	254	133.0	76.0	22.91	75	76
12	43	0	247	131.0	88.0	27.64	72	61
13	46	15	294	142.0	94.0	26.31	98	64
14	41	0	332	124.0	88.0	31.31	65	84
15	39	9	226	114.0	64.0	22.35	85	5500
16	38	20	221	140.0	90.0	21.35	95	70
17	48	10	232	138.0	90.0	22.37	64	72
18	46	20	291	112.0	78.0	23.38	80	89
19	38	5	195	122.0	84.5	23.24	75	78
20	41	0	195	139.0	88.0	26.88	85	65
21	42	30	190	108.0	70.5	21.59	72	85
22	43	0	185	123.5	77.5	29.89	70	NA
23	52	0	234	148.0	78.0	34.17	70	113
24	52	20	215	132.0	82.0	25.11	71	75
25	44	30	270	137.5	90.0	21.96	75	83

fram%>%select(age,cigsPerDay,totChol,sysBP,diaBP,BMI,heartRate,glucose)->conti ; View(conti)

*결과값이 너무 많아서 뒷부분은 생략함.

```
> round(cor(conti2),2)
           age cigsPerDay totChol sysBP diaBP   BMI heartRate glucose
age       1.00      -0.19    0.26  0.39  0.20  0.14     -0.01    0.01
cigsPerDay -0.19     1.00   -0.03 -0.09 -0.06 -0.09      0.07   -0.02
totChol    0.26     -0.03    1.00  0.22  0.17  0.12      0.09    0.01
sysBP      0.39     -0.09    0.22  1.00  0.78  0.33      0.18    0.00
diaBP      0.20     -0.06    0.17  0.78  1.00  0.38      0.18   -0.01
BMI        0.14     -0.09    0.12  0.33  0.38  1.00      0.07    0.01
heartRate -0.01      0.07    0.09  0.18  0.18  0.07      1.00    0.04
glucose    0.01     -0.02    0.01  0.00 -0.01  0.01      0.04    1.00
```

library(dplyr); conti2<-na.omit(conti); round(cor(conti2),2)



round(cor(conti2),2)->a

corrplot(a,method="circle",tl.col="black",tl.srt=30)

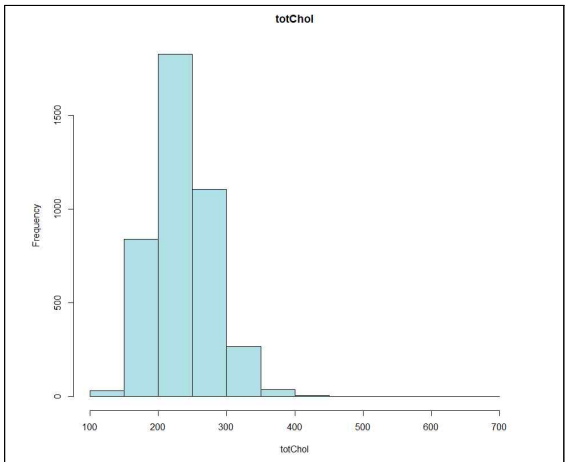
상관계수를 숫자로 표현하였고 이를 더 직관적으로 판단할 수 있도록 시각화하였다. 푸른 계열로 갈수록 양의 상관관계를 뜻하며, 반대로 붉은 계열은 음의 상관관계를 뜻한다. 원의 크기가 클수록 상관관계가 강하다는 것이다.

한편 이를 통해 연속형 변수간 상관관계를 따져볼 수 있다. 이를 이용해서 연속형 변수 안에 있는 결측값을 추론해볼 것이다.

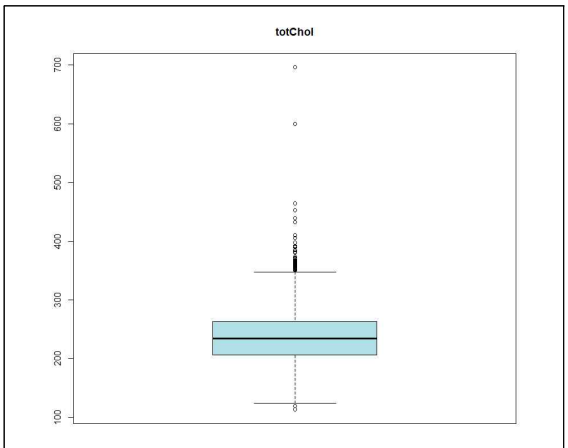
#totChol

totChol은 콜레스트롤 수치를 나타내는 변수로 50개 정도의 결측값들을 가지고 있다. 하지만 결측값들은 다른 변수와의 연관성을 찾을 수 없어 삭제 처리하였다. 이를 바탕으로 연속형 변수에 맞게 히스토그램과 상자그림을 그렸다.

콜레스트롤의 평균은 236.7mg/dL 이며, 정상 콜로스트롤 범위가 0~240mg/dL 인 것을 따져보면 정상적인 수치임을 알 수 있다. 그래프의 형태는 우향 왜곡 분포로 왼쪽에 많이 치우쳐져 있다.



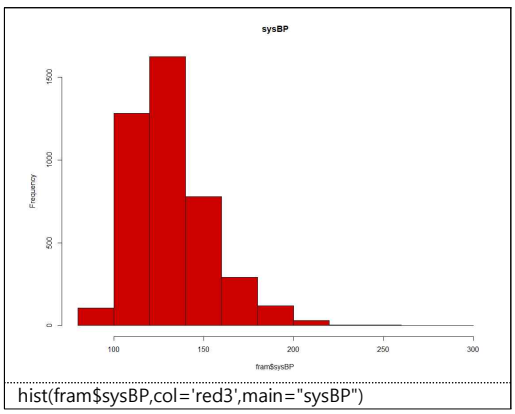

```
fram%>%filter(!is.na(totChol))->fram
hist(fram$totChol,col="powder  blue",main="totChol",xlab='totChol')
```



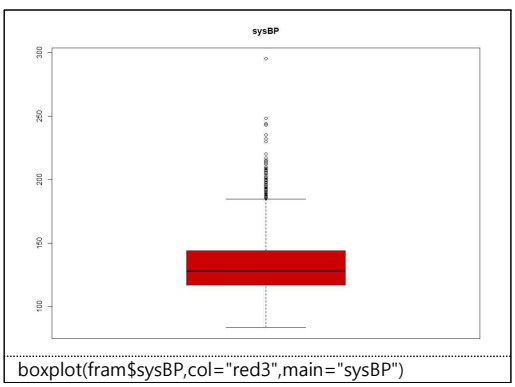

```
boxplot(fram$totChol,col="powder  blue",main="totChol")
```

#sysBP

sysBP는 수축기 혈압을 나타내는 변수로, 특별한 오류가 없었기 때문에 바로 연속형 변수의 그래프인 히스토그램과 상자그림을 그려 파악하였다. 그래프의 전체적인 분포는 우향 왜곡 분포이며 왼쪽에 값이 몰려있음을 알 수 있다. 수축기 혈압의 평균은 132.2mmHg 이며, 정상인 평균 수치가 120mmHg 미만인 것을 고려했을 때 다소 높다는 것을 알 수 있다.




```
hist(fram$sysBP,col='red3',main="sysBP")
```




```
boxplot(fram$sysBP,col="red3",main="sysBP")
```

#diaBP

diaBP는 이완기 혈압을 나타내는 변수로, 30개의 결측값을 가지고 있다. 이 변수의 경우 상관계수 그래프를 보면 알 수 있듯이 sysBP와 가장 관련 되어있다. 두 변수 간 상관계수의 경우 0.78로 상당히 높은 양의 상관관계이다.

따라서 diaBP의 결측값은 sysBP와의 회귀분석식으로 추정하여 값을 수정하였다. sysBP의 값에 따른 diaBP의 값이므로 독립변수는 sysBP이고 종속변수는 diaBP이다. 이를 바탕으로 회귀분석을 하였고 $\hat{y}= 0.4247x + 26.6885$ 라는 회귀식을 도출하였다. 이후 도출된 회귀분석식을 토대로 diaBP 변수의 결측값을 수정하였다.

수정 전

115.0	85.5
150.0	85.0
134.0	82.5
147.0	74.0
124.5	NA
153.5	102.0
160.0	98.0
153.0	101.0
111.0	73.0

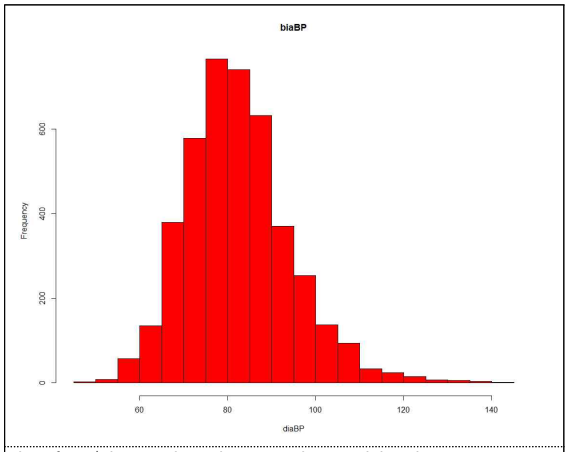
```
fram[is.na(fram$diaBP),]
```

수정 후

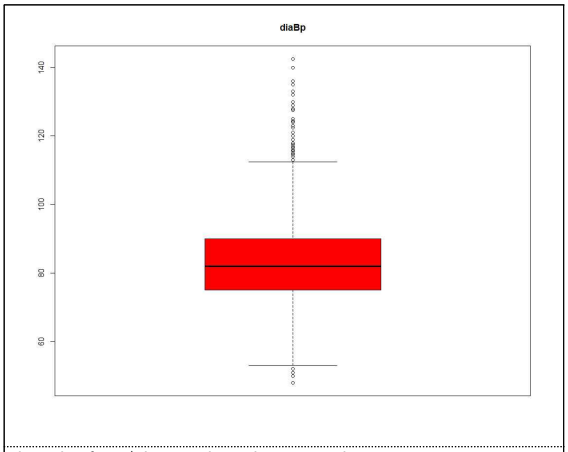
115.0	85.50000
150.0	85.00000
134.0	82.50000
147.0	74.00000
124.5	79.56365
153.5	102.00000
160.0	98.00000
153.0	101.00000
111.0	73.00000

```
lm(conti2$diaBP~conti2$sysBP)
#y = 0.4247x + 26.6885 회귀대체 (NA 대체값)
for(i in 1:4240){
  if(is.na(fram$diaBP[i]) == TRUE){
    fram$diaBP[i] <- 0.4247*fram$sysBP[i] + 26.6885
  }
}
fram
sum(is.na(fram$diaBP))
View(fram)
```

수정한 내용을 바탕으로 biaBP의 그래프를 그렸다. 연속형 분포에 맞게 히스토그램과 상자그림을 그렸으며, 그래프의 전체적인 양상이 정규분포와 비슷하게 대칭 형태를 이루고 있음을 확인할 수 있다. 또한 이완기 혈압의 평균은 82.9mmHg이며, 정상인 평균 수치가 80mmHg 임으로 고려하면 정상적인 수치임을 확인할 수 있다.




```
hist(fram$diaBP,col='red1',main='biaBP',xlab='diaBP')
```




```
boxplot(fram$diaBP,col='red1',main='diaBp')
```

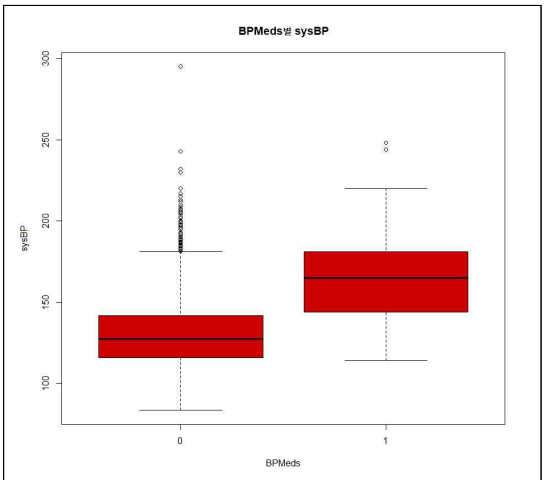
#sysBP와 수정한 diaBP를 바탕으로, BPMeds 판단

정상인 혈압은 120/80이다.

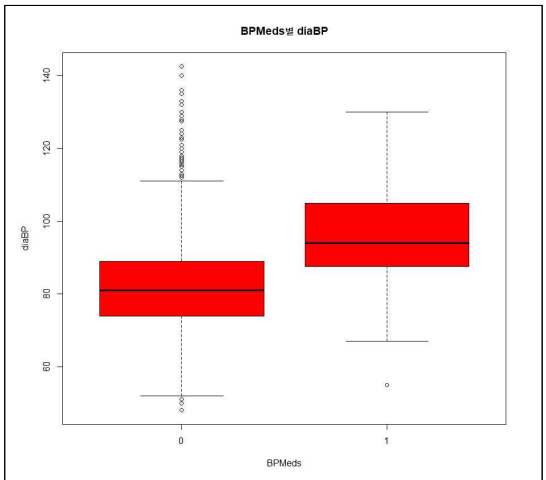
우선 BPMeds=1, 즉 혈압약을 복용하는 이들만 모아놓은 테이블 BPM 생성하였다. 이후 BPM 테이블의 분포 양상 파악하였다. 분석하기 전, 아마도 고혈압 환자들이 약을 복용할 것이라 예측하였는데, 예상과 맞게 수축기 혈압이 120 이상이고 이완기 혈압이 80이상인 조건이 모두 맞는 이들만 약을 복용한다는 점을 알아내었다. 하지만 그렇지 않은 경우도 다반수였기에 확정할 순 없었다. 따라서 우선 결측치를 가진 행들은 삭제하였다. 확신이 없기에 값을 수정하진 않았다.

*결측치 제거

```
> fram%>%filter(!is.na(BPMeds))->fram
> sum(is.na(fram$BPMeds))
[1] 0
```



boxplot(sysBP~BPMeds,data=fram)



boxplot(diaBP~BPMeds,data=fram)

정제된 테이블을 가지고 BPMeds에 따른 sysBP와 diaBP의 그래프를 그려보았다. 상자그림을 그려 범주별로 나누어 비교해보니 두 그룹 사이의 차이가 더 명확해짐을 알 수 있다. 확실히 혈압약을 복용하는 환자들의 경우 평균 혈압이 높게 측정되고 있다.

한편 혈압약 복용하는 이들의 공통점이 있는데, 바로 prevalentHyp=1이라는 점이다. 테이블에서의 혈압약을 복용하는 모든 이들이 우울증상이 발현되고 있다. 실제 일부 혈압약은 우울증을 유발한다는 연구결과가 있다.[2)]

BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP
1	0	1	0	332	124.0	88.0
1	0	1	0	NA	148.0	92.0
1	0	1	1	311	206.0	92.0
1	0	1	0	NA	125.0	80.0
1	0	1	0	254	191.0	124.5
1	0	1	0	244	139.0	86.0
1	1	1	0	252	189.0	110.0
1	0	1	0	283	130.0	80.0
1	0	1	0	368	204.0	94.0
1	0	1	0	250	144.0	98.0
1	0	1	0	201	156.0	93.0

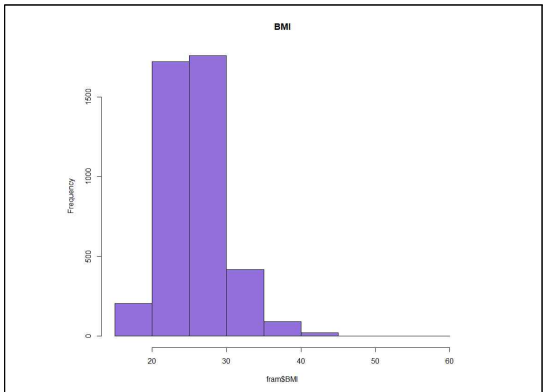
```
fram%>%filter(BPMeds==1)->BPM
View(BPM)
```

2) https://www.pharmnews.com/news/articleView.html?idxno=84031

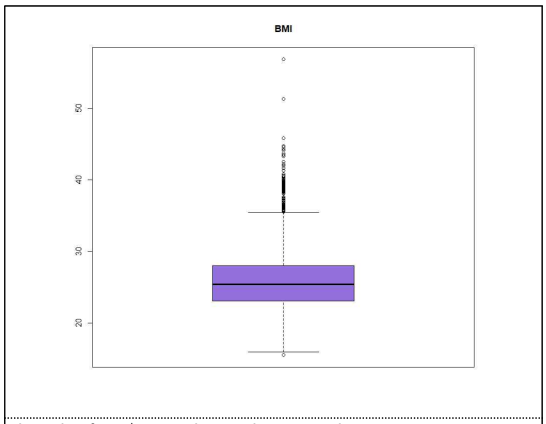
#BMI

BMI는 체질량 지수를 나타내는 변수이다. 이 변수의 경우 결측값이 19개였는데 다른 변수와의 연관성은 찾을 수 없었다. 따라서 결측값 19개는 삭제하였으며, 전체 4240개의 데이터 중 19개의 값을 삭제한 것이므로, 삭제에 따른 데이터 변화는 크지 않았다.

이후 정리된 데이터를 바탕으로 연속형 변수에 맞게 히스토그램과 상자그림을 그려 분포 양상을 파악했다. 대부분이 20~30 사이에 분포하고 있음을 확인하였는데, 실제로 정상 체질량 지수는 20~24이다. 따라서 정상 범주와 경도비만 범주에 많은 사람이 몰려있다고 판단하였다.



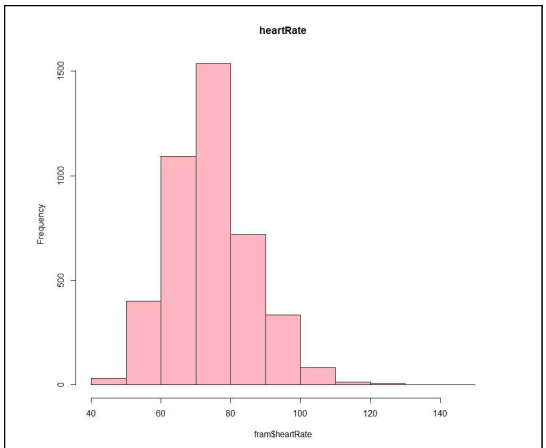

```
fram%>%filter(!is.na(BMI))->fram #결측치 제거
hist(fram$BMI,col="medium purple",main="BMI")
```



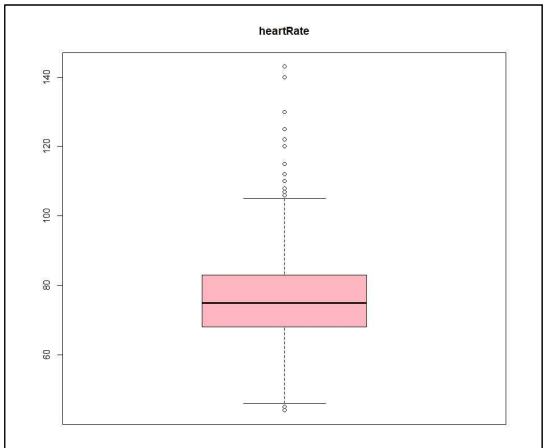

```
boxplot(fram$BMI,col="medium purple",main="BMI")
```

#heartRate

heartRate는 분당 심박수를 나타내는 변수이다. 결측값이 존재하였으나 1개뿐이어서 수정하지 않고 삭제 처리하였다. 이를 바탕으로 연속형 변수에 맞게 히스토그램과 상자그림을 그렸다. 그래프를 보면 대부분의 분포가 68~83회 사이에 있음을 알 수 있다. 실제 정상 심박수 범위는 60~100회 사이로, 자료 속 대부분의 사람이 정상범주에 속하고 있음을 알 수 있다.




```
fram%>%filter(!is.na(heartRate))->fram #결측치 제거
hist(fram$heartRate,col="light pink",main="heartRate")
```




```
boxplot(fram$heartRate,col="light pink",main="heartRate")
```

#glucose

우선, glucose의 임상적 의의를 살펴보자. 주로 혈당을 나타내는 수치로 당뇨병 진단을 위해 검사하는 수치이다. 특히 당뇨병 진단 시 이용되는 혈당은 공복혈당이므로, 이 변수도 공복혈당 수치일 것이라 가정하였다. glucose의 경우 정상 범위는 80~130mg/dl 정도 이다. 당뇨병으로 진단되고 치료가 되는 기준치는 보통 126mg/dl 정도이다.

(참고) 당뇨병 기준 자료

[혈당수치 조견표] (세계보건기구 자료)

진단	공복 혈당수치	식후2시간 혈당수치
정상	70~110mg/dl	70~140mg/dl
당뇨 전단계, 예비당뇨 (공복혈당장애)	110~125mg/dl	70~140mg/dl
당뇨 전단계, 예비당뇨 (내당능장애)	110~125mg/dl	140~200mg/dl
당뇨판정	126mg/dl 이상	200mg/dl 이상

* **공복의 정의** : 식후 12시간

따라서 glucose는 공복 혈당 수치라고 생각하고 분석을 진행하였다. 한편 glucose의 경우 결측값이 383개나 존재한다. 또한 5500 이라는 큰 수의 이상점을 여러 개 가지고 있다.

우선 glucose에서 이상치에 해당되는 값들부터 처리하였다. 1000이상의 수치는 이상값으로 보았으며 검색결과 모두 glucose=5500에 해당되는 값들이었다. 이를 glucose=55인 값들과 비교해보았는데 다른 변수들이 모두 유사한 양상을 보임을 확인하였고, 5500인 값들을 삭제하기 보단 55로 수정하였다.

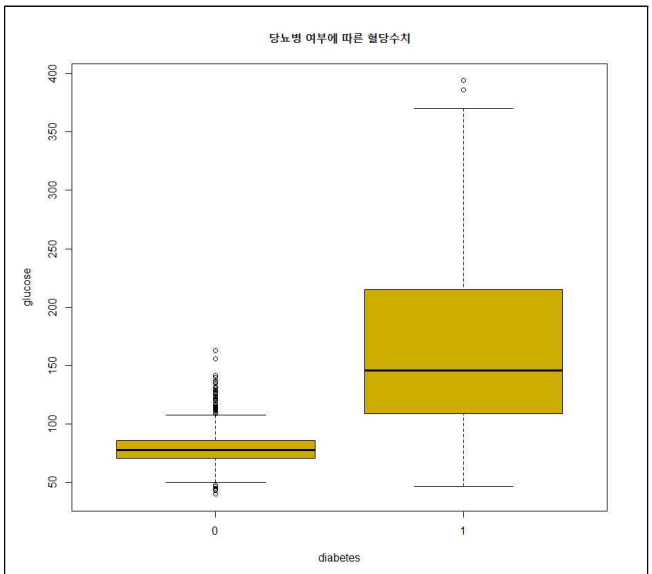
```
> fram%>%filter(glucose>1000)
  male age education currentSmoker cigsPerDay BPMeds prevalentStroke prevalentHyp diabetes totChol sysBP    diaBP   BMI heartRate glucose TenYearCHD
1    0  39         2             1          9      0               0            0        0      226 114.0 64.00000 22.35        85    5500          0
2    0  60         1             0          0      0               0            0        0      260 110.0 72.50000 26.59        65    5500          0
3    0  63         4             0          0      0               0            0        0      248 164.5 96.55165 29.35        70    5500          0
4    0  43         3             0          0      0               0            0        0      263 115.0 82.50000 25.91       105    5500          0
```

```
> fram%>%filter(glucose==55)
   male age education currentSmoker cigsPerDay BPMeds prevalentStroke prevalentHyp diabetes totChol sysBP diaBP   BMI heartRate glucose TenYearCHD
1     0  54         1             1          9      0               0            0        1      266 114.0  76.0 17.61        88      55          0
2     0  45         2             1         20      0               0            0        0      285 116.0  87.0 23.85        65      55          1
3     0  41         1             1          4      0               0            0        0      176 113.0  75.0 22.29        80      55          0
4     1  48         1             0          0      0               0            0        0      270 131.0  88.0 27.13        63      55          0
5     1  40         4             1         20      0               0            0        0      229 137.0  85.0 35.20        66      55          0
6     1  38         1             0          0      0               0            0        0      221 130.0  87.0 26.43        72      55          0
7     1  43         1             0          0      1               0            1        0      210 181.0  97.5 21.83        75      55          0
8     0  43         3             1          1      0               0            0        0      185 125.0  84.0 23.18        75      55          0
9     1  41         1             1         40      0               0            0        0      260 137.5  80.0 26.89        75      55          1
10    0  51         2             1          5      0               0            0        0      315 119.0  75.0 25.79        75      55          0
11    0  42         3             0          0      0               0            0        0      215 111.0  72.0 25.38        77      55          0
12    1  38         4             1         15      0               0            0        0      248 110.0  61.0 22.17        85      55          1
13    1  42         2             1         30      0               0            0        0      209 130.0  86.0 24.01        60      55          0
```

▼ glucose=5500 → glucose=55

```
> fram$glucose[fram$glucose==5500]<-55
> fram%>%filter(glucose==55)
   male age education currentSmoker cigsPerDay BPMeds prevalentStroke prevalentHyp diabetes totChol sysBP    diaBP   BMI heartRate glucose TenYearCHD
1     0  39         2             1          9      0               0            0        0      226 114.0 64.00000 22.35        85      55          0
2     0  60         1             0          0      0               0            0        0      260 110.0 72.50000 26.59        65      55          0
3     0  54         1             1          9      0               0            0        1      266 114.0 76.00000 17.61        88      55          0
4     0  63         4             0          0      0               0            0        0      248 164.5 96.55165 29.35        70      55          0
5     0  43         3             0          0      0               0            0        0      263 115.0 82.50000 25.91       105      55          0
6     0  45         2             1         20      0               0            0        0      285 116.0 87.00000 23.85        65      55          1
7     0  41         1             1          4      0               0            0        0      176 113.0 75.00000 22.29        80      55          0
8     1  48         1             0          0      0               0            0        0      270 131.0 88.00000 27.13        63      55          0
9     1  40         4             1         20      0               0            0        0      229 137.0 85.00000 35.20        66      55          0
10    1  38         1             0          0      0               0            0        0      221 130.0 87.00000 26.43        72      55          0
11    1  43         1             0          0      1               0            1        0      210 181.0 97.50000 21.83        75      55          0
12    0  43         3             1          1      0               0            0        0      185 125.0 84.00000 23.18        75      55          0
13    1  41         1             1         40      0               0            0        0      260 137.5 80.00000 26.89        75      55          1
14    0  51         2             1          5      0               0            0        0      315 119.0 75.00000 25.79        75      55          0
15    0  42         3             0          0      0               0            0        0      215 111.0 72.00000 25.38        77      55          0
16    1  38         4             1         15      0               0            0        0      248 110.0 61.00000 22.17        85      55          1
17    1  42         2             1         30      0               0            0        0      209 130.0 86.00000 24.01        60      55          0
```

한편 앞서 언급한 바와 같이 glucose는 공복혈당 수치라고 생각했을 때, 당뇨병 진단 시 이용되는 수치이다. 따라서 glucose와 diabetes가 연관되어있을 것이라고 보았으며, 당뇨병 질환이 있는 경우 glucose의 수치가 정상인보다 높게 측정될 것이라고 예측하였다. 이를 바탕으로 diabetes의 범주에 따른 glucose 수치를 상자그림으로 표현하였고, 이는 다음과 같다.



boxplot(glucose~diabetes,data=fram,col='gold3',main='당뇨병 여부에 따른 혈당수치')

예상한 바와 동일하게, 당뇨병 기저질환이 있는 그룹의 glucose 수치가 정상인보다 더 높게 잡힘을 확연히 알 수 있다. 따라서 diabetes=1인 경우는 glucose의 수치도 높게 측정되었을 것이라 판단하였다.

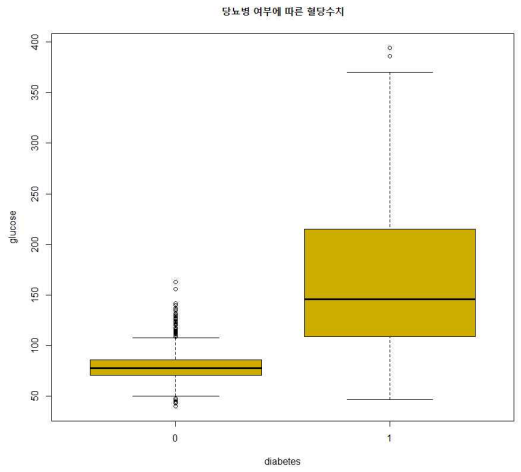
한편 glucose의 값이 결측값인 경우의 행들을 모두 검색해보았는데, 다음과 같이 모두 diabetes=0 인 경우임을 알 수 있다. 삭제하기에는 400개가 넘는 데이터들이여서 수정하는 방향으로 갔다.

```
> fram[is.na(fram$glucose),]
    male age education currentSmoker cigsPerDay BPMeds prevalentStroke prevalentHyp diabetes totChol sysBP diaBP   BMI heartRate glucose TenYearCHD
22     0  43         1             0          0      0               0            0        0      185 123.5  77.5 29.89        70      NA          0
53     0  39         2             1         20      0               0            0        0      209 115.0  75.0 22.54        90      NA          0
108    1  51         3             0          0      0               0            1        0      214 145.0  92.5 26.09        70      NA          0
111    0  41         1             0          0      0               0            1        0      265 136.0  98.0 42.15        90      NA          0
128    1  43         4             1         18      0               0            0        0      222 109.5  69.0 25.50        75      NA          0
198    1  41         1             1         40      0               0            0        0      239 119.5  70.0 29.79        70      NA          0
206    1  61         1             0          0      0               0            0        0      239 143.0  80.0 25.74        48      NA          0
210    1  45         1             1         43      0               0            1        0      191 139.5  75.0 22.30        77      NA          0
240    0  61         3             0          0      0               0            1        0      189 133.0  83.0 22.82        87      NA          1
243    1  38         1             1         20      0               0            0        0      268 117.0  83.0 33.61        72      NA          0
256    1  64         1             0          0      0               0            1        0      217 147.0  87.0 29.73        77      NA          0
266    0  43         4             0          0      0               0            0        0      213 100.0  70.0 20.06        68      NA          0
272    0  45         3             0          0      0               0            0        0      129 109.0  69.0 22.36        75      NA          0
275    0  35         2             0          0      0               0            0        0      208 122.5  72.5 22.00        65      NA          0
288    1  51         3             0          0      0               0            0        0      245 124.0  69.0 21.52        85      NA          0
293    0  48         3             0          0      0               0            0        0      230 129.0  84.5 24.73        78      NA          0
294    0  55         1             0          0      0               0            0        0      220 117.5  84.0 26.20        90      NA          0
```

*결과값이 너무 많아서 뒷부분은 생략함.

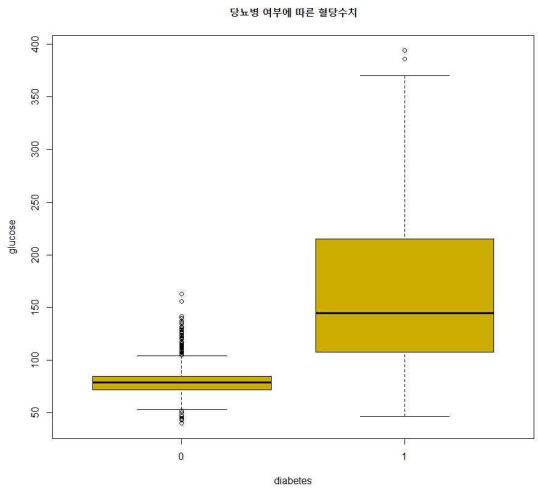
이 경우 diabetes=0 이므로 당뇨병이 없는 경우이다. 따라서 glucose는 정상인 수치에 있을 것이라 가정하였고, 이를 바탕으로 결측값을 평균값으로 대체하였다. 당뇨병이 없는 그룹의 평균 glucose이 79이므로 결측값은 79로 수정하였다. 이를 반영하여 diabetes 그룹별 glucose의 분포를 다시 그려보았다. 다음과 같이 큰 변화 없음을 확인하였다.

수정 전



boxplot(glucose~diabetes,data=fram,col='gold3',main='당뇨병 여부에 따른 혈당수치')

수정 후




```
fram[fram$diabetes==0,]->nondia
nondia.mean<-mean(nondia$glucose,na.rm=TRUE)
round(nondia.mean)
fram$glucose[is.na(fram$glucose)]<-79
sum(is.na(fram$glucose))
boxplot(glucose~diabetes,data=fram,col='gold3',main='당뇨병 여부에 따른 혈당수치')
```

#TenYearCHD

TenYearCHD은 십 년 후 관상동맥질환 발병 여부에 대한 변수이다. TenYearCHD의 경우 0,1 이외의 '.' 값이 3개 존재한다. 따라서 '.'값을 수정할 수 있는지 보기 위해, TenYearCHD=. 인 경우의 행들을 검색하였고, TenYearCHD의 중요 포인트가 될 수 있는 TenYearCHD=1 인 경우의 행들을 검색하여 두 경우를 비교해보았다.

TenYearCHD=1 이 중요하다고 생각한 이유는, TenYearCHD의 경우 10년 후 관상동맥질환이 발병하느냐를 나타내는 변수이므로, 10년 후 관상동맥질환이 발병하는 경우엔 나머지 변수가 어떤 값들을 보이고 있었는지를 파악하는 것이 중요하다고 생각했기 때문이다. 따라서 '.' 인 경우의 나머지 변수 형태와 1일 때 나머지 변수의 형태를 비교해보았다.

```
> fram%>%filter(TenYearCHD==".")
  male age education currentSmoker cigsPerDay BPMeds prevalentStroke prevalentHyp diabetes totChol sysBP diaBP   BMI heartRate glucose TenYearCHD
1    0  36         1             1          3      0                0            0        0      135   108    74 22.53        73      75          .
2    0  57         1             0          0      0                0            0        0      197    96    64 18.59        60      77          .
3    0  49         3             1          3      0                0            0        0      247   121    82 29.07        72      69          .
```

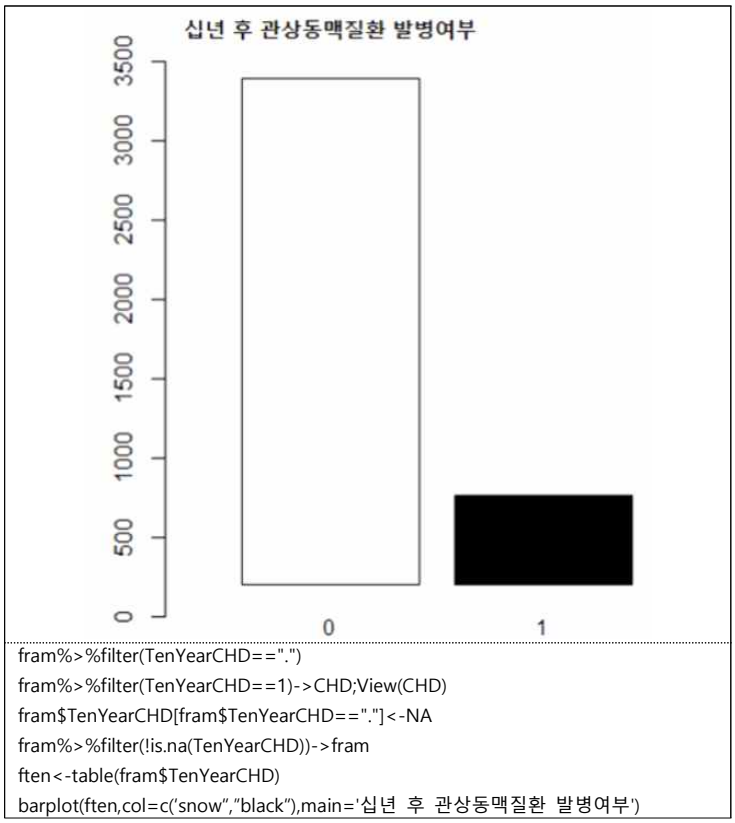
	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
351	1	66	2	0	0	0	0	1	0	124	138.0	91.0000	32.33	75	96	1
399	1	51	1	1	7	0	0	1	0	133	138.0	78.0000	16.98	80	65	1
112	1	47	1	1	30	0	0	0	0	143	114.0	79.0000	26.59	69	72	1
285	0	40	1	0	0	0	0	0	0	144	122.5	80.0000	27.46	72	123	1
502	1	58	2	0	0	0	0	1	0	149	98.0	60.0000	24.73	105	71	1
309	0	58	1	1	20	1	0	1	0	156	170.0	98.0000	28.88	72	79	1
520	1	68	1	1	15	0	0	0	0	157	106.0	48.0000	26.73	65	65	1
257	1	62	1	1	10	0	0	0	0	157	134.0	84.0000	25.95	105	76	1
489	1	44	1	1	40	0	0	0	0	158	150.5	87.0000	21.44	75	98	1
159	1	57	1	1	20	0	0	1	0	158	154.0	100.0000	24.07	92	70	1

*결과값이 너무 많아서 뒷부분은 생략함. (fram%>%filter(TenYearCHD==1))

각 변수들의 분포를 따져보아도 TenYearCHD의 발병 여부는 한 눈에 판단하기 쉽지 않다. 기저질환이 없어도, 평소 혈압이 정상이어도, 혈당이 정상수치일지라도 십 년 뒤 관상동맥

질환이 발병하는 경우가 많이 나왔다. 따라서 우선은 TenYearCHD=. 인 경우는 결측값으로 변환시켜 삭제하였다.

정제된 데이터 테이블을 가지고 TenYearCHD에 대한 그래프를 그려보았다. 범주형 변수에 맞게 막대그래프를 그려 그래프의 형태를 보았다. 그래프는 십년 후 관상동맥질환이 발병되는 경우가 적었지만 그럼에도 불구하고 613건이나 존재한다는 것을 알 수 있다.




```
fram%>%filter(TenYearCHD==".")
fram%>%filter(TenYearCHD==1)->CHD;View(CHD)
fram$TenYearCHD[fram$TenYearCHD=="."]<-NA
fram%>%filter(!is.na(TenYearCHD))->fram
ften<-table(fram$TenYearCHD)
barplot(ften,col=c('snow","black"),main='십년 후 관상동맥질환 발병여부')
```

```
   0    1
3503  613
```

#Target Variable (TenYearCHD)와 변수 사이의 관계

지금까지 각 변수 별 형태를 살펴보았다. 결국 타겟변수는 TenYearCHD이므로 각 변수와 TenYearCHD 간의 관계를 살펴볼 것이다.

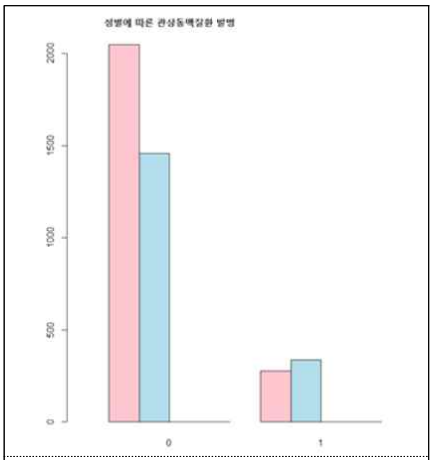
#male 과 TenYearCHD

다음은 성별에 따른 십년 후 관상동맥질환 여부에 대한 그래프이다. 가장 먼저 범주형 변수 간 비교를 위해 부분합 테이블을 만들었고 이를 바탕으로 성별에 따른 관상동맥질환 발병 여부를 그린 막대그래프를 그렸다.

아래 그래프의 분홍색은 여성이며 하늘색은 남성을 뜻한다. 한눈에 보아도 각 범주 모두 여성의 수가 더 많았다. 하지만 십년 후 관상동맥질환 발병 횟수는 남성이 여성보다 더 많음을 볼 수 있다. 따라서 발병 비율로 따져보면 **남성이 여성보다 십년 후 관상동맥질환 발병률이 더 높다고 할 수 있다.**

	.	0	1
0	0	2047	277
1	0	1456	336
F	0	0	0
M	0	0	0

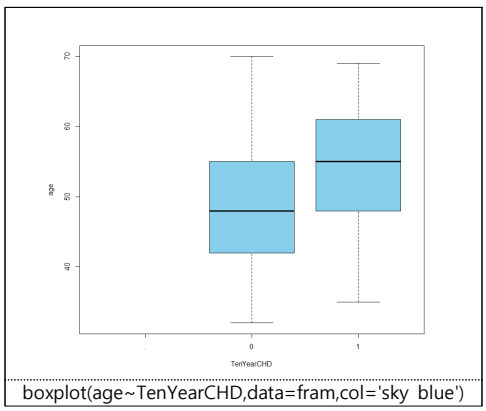
```
table(fram$male,fram$TenYearCHD,useNA='ifany')->maletg
```




```
barplot(maletg,main='성별에 따른 관상동맥질환 발병',beside=TRUE,col=c('pink','light blue'))
```

#age 와 TenYearCHD

나이에 따른 TenYearCHD는 아래의 그래프와 같다. 십년 후 관상동맥질환이 발병되는 경우의 평균 나이가 발병되지 않았을 때보다 높은 것을 알 수 있다. **즉 나이가 들어감에 따라 관상동맥 질환에 취약하다고 해석할 수도 있다.**



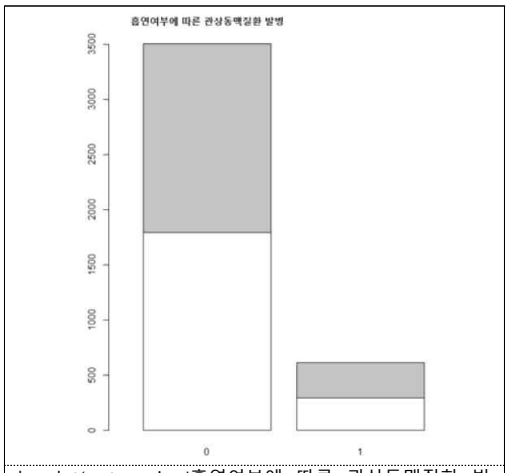

```
boxplot(age~TenYearCHD,data=fram,col='sky blue')
```

#current Smoker 와 TenYearCHD

현재 흡연자와 십년 뒤 관상동맥질환 발병 여부의 관계는 다음과 같다. 회색 범위는 흡연자이고 하얀색 범위는 비흡연자의 범위이다. 그래프를 보면 한 눈에 알아보기 힘들만큼 흡연여부에 따른 관상동맥질환 발병이 얼마나 차이 나는지 알 수 없다. 확실한 점은 질병의 발병의 경우가 더 적은 비율을 가지고 있다. 그러나 **흡연의 여부와는 관련 짓기 힘들다.**

```
  .    0   1
0 0 1791 293
1 0 1712 320
```

```
table(fram$currentSmoker,fram$TenYearCHD,useNA='ifany')->smtg
```

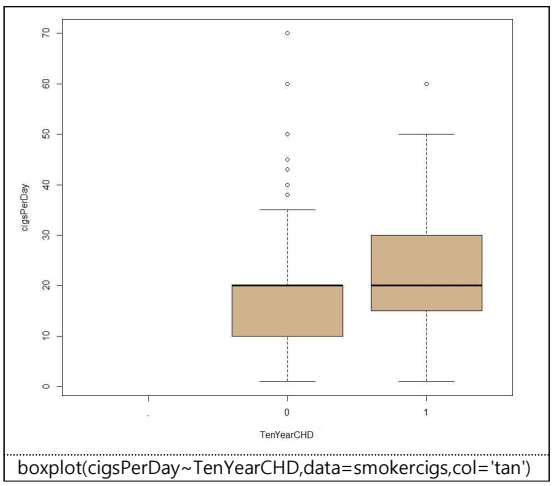



```
barplot(smtg,main='흡연여부에 따른 관상동맥질환 발병',besides=TRUE,col=c('snow','gray'))
```

#cigs per day 와 TenYearCHD

흡연량과 십년 뒤 관상동맥질환 발병 여부를 살펴보기에 앞서, 흡연량과의 비교를 위해 전제는 흡연자들인 경우로 두었다. 비흡연자를 넣게 될 경우 흡연량 평균에 변화가 생겨 정확한 판단을 하기 힘들 것이라 판단하였다. (평균이 더 아래로 잡히는 경우) 평균에 대한 변화가 정확한 판단을 흐릴거라고 생각한 이유는 앞서 분석한 current Smoker와 TenYearCHD 관계에서 흡연의 여부는 질병 발생과 무관하다고 판단하였기 때문이다.

따라서 두 변수사이의 관계에 앞서 전제는 흡연자들의 흡연량으로 두었다. 이에 따라 그래프를 그려보면 다음과 같다. 그래프를 보면 발병하는 경우의 흡연량 분포가 더 넓었다. 하지만 흡연량 중앙값은 비슷하게 잡히는 것으로 보아 **섣불리 흡연량의 정도도 관상동맥질환 발병 여부와 관련있다고 판단하기엔 쉽지않다고 생각했다.**



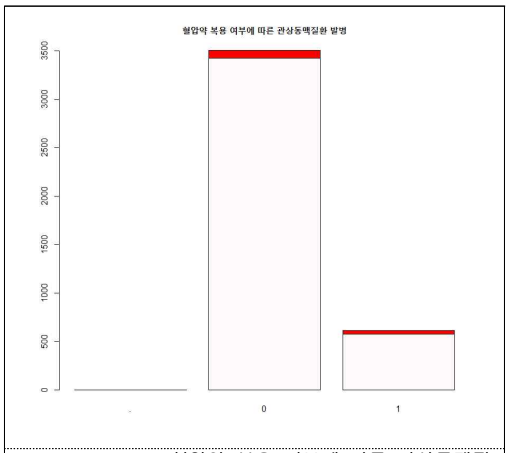

```
boxplot(cigsPerDay~TenYearCHD,data=smokercigs,col='tan')
```

#BMPeds와 TenYearCHD

혈압약 복용 여부와 십년 뒤 관상동맥질환 발병 여부의 관계는 아래와 같다. 붉은 범위가 바로 혈압량을 복용한 경우인데, 모두 각 범주에서 매우 작은 비율을 차지하고 있다. **따라서 혈압량 복용여부는 질병 발병 여부와 큰 관련이 있지 않다고 판단하였다.**

```
     .    0    1
0    0 3422  574
1    0   81   39
```

table(fram$BPMeds,fram$TenYearCHD,useNA='ifany')->bp;bp



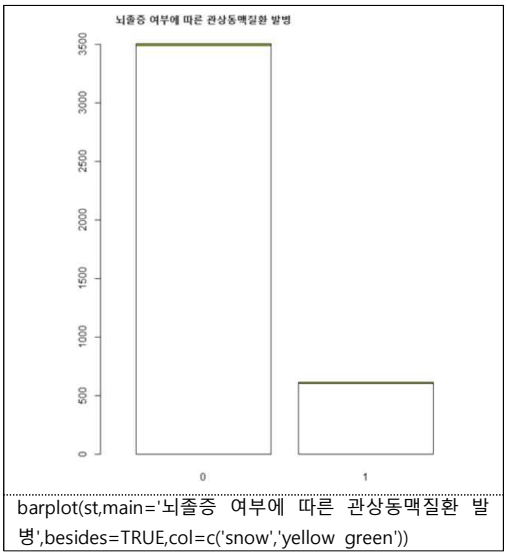
barplot(bp,main='혈압약 복용 여부에 따른 관상동맥질환 발병',besides=TRUE,col=c('snow','red1'))

#prevalentStroke와 TenYearCHD

뇌졸증 여부와 십년 뒤 관상동맥질환 발병 여부의 관계는 아래와 같다. 녹색 범위가 바로 뇌졸중 여부인데, 모두 각 범주에서 매우 작은 비율을 차지하고 있다. **따라서 뇌졸중 여부는 질병 발병 여부와 큰 관련이 있지 않다고 판단하였다.**

```
     .    0    1
0    0 3489  605
1    0   14    8
```

table(fram$prevalentStroke,fram$TenYearCHD,useNA='ifany')->st



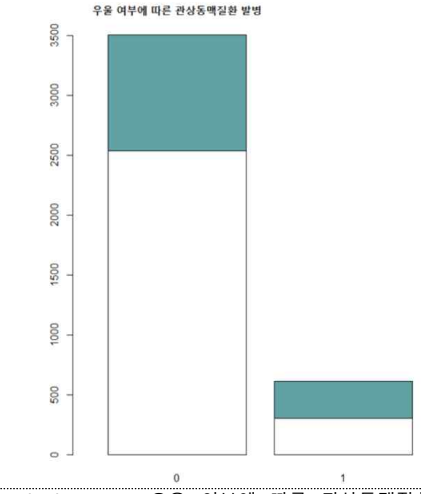
barplot(st,main='뇌졸증 여부에 따른 관상동맥질환 발병',besides=TRUE,col=c('snow','yellow green'))

#prevalentHyp와 TenYearCHD

우울의 여부와 십년 뒤 관상동맥질환 발병 여부의 관계는 아래와 같다. 녹색 범위가 바로 우울 증상 여부인데, 질병이 발병되지 않은 경우 적은 비율을 차지하고 있지만 질병이 발병되는 경우 말이 달라진다. 우울을 느낀 상당수의 인원들이 십년 뒤 관상동맥질환을 겪게 되었으며 이는 TenYearCHD에서 절반 가량의 비율을 차지하고 있음을 알 수 있다.

```
     .    0    1
0    0 2538  306
1    0  965  307
```

```
table(fram$prevalentHyp,fram$TenYearCHD,useNA='ifany')->hyp
```



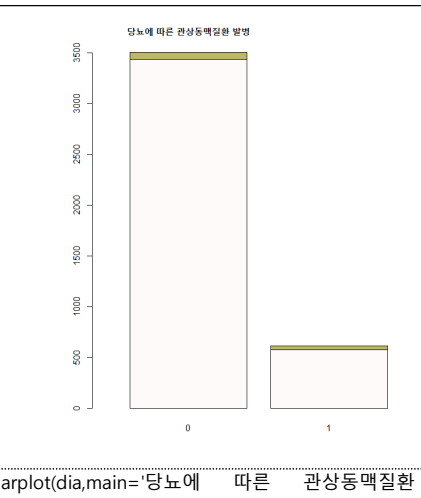

```
barplot(hyp,main='우울 여부에 따른 관상동맥질환 발병',besides=TRUE,col=c('snow','cadet blue'))
```

#diabetes와 TenYearCHD

당뇨의 여부와 십년 뒤 관상동맥질환 발병 여부의 관계는 아래와 같다. 녹색 범위가 바로 당뇨증 여부 범위인데 크게 관계가 없음을 알 수 있다.

```
     .    0    1
0    0 3435  577
1    0   68   36
```

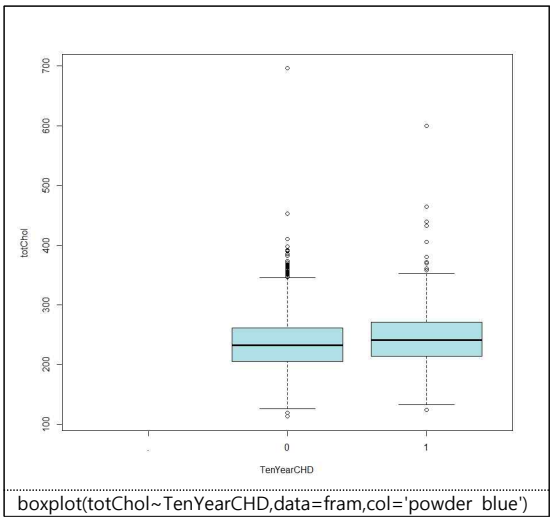
```
table(fram$diabetes,fram$TenYearCHD,useNA='ifany')->dia
```




```
barplot(dia,main='당뇨에 따른 관상동맥질환 발병',besides=TRUE,col=c('snow','dark khaki'))
```

#totChol와 TenYearCHD

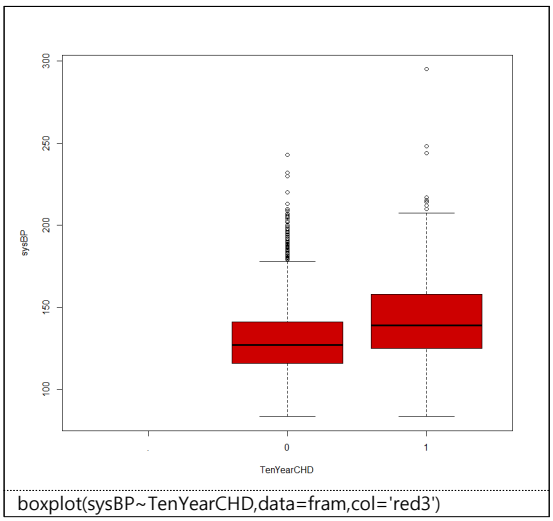
콜레스트롤의 수치에 따른 십년 뒤 관상동맥질환 발병 여부는 관계가 없다고 판단하였다.




```
boxplot(totChol~TenYearCHD,data=fram,col='powder  blue')
```

#sysBP와 TenYearCHD

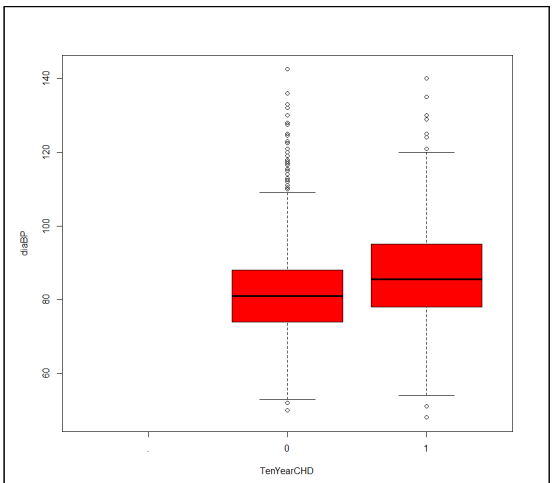
수축기 혈압에 따른 십년 뒤 관상동맥질환 발병 여부의 그래프는 다음과 같다. 관상동맥질환이 발병되는 경우의 혈압이 더 높게 측정됨을 알 수 있다. **따라서 고혈압일수록 관상동맥질환이 발병되기 쉽다고 판단할 수 있다.**




```
boxplot(sysBP~TenYearCHD,data=fram,col='red3')
```

#diaBP와 TenYearCHD

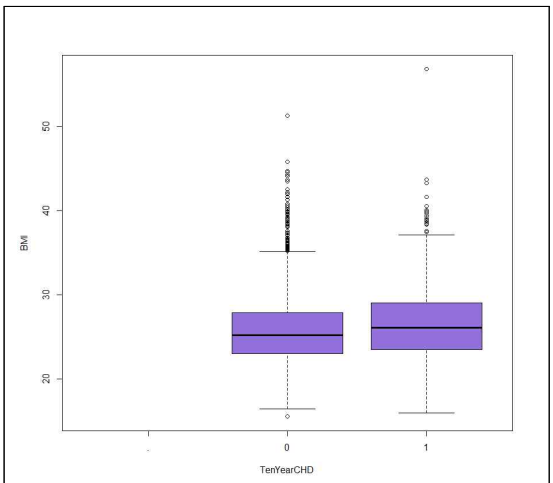
이완기 혈압에 따른 관상동맥질환 발병 여부의 그래프는 다음과 같다. 수축기 혈압과 마찬가지로 관상동맥질환이 발병된 경우의 이완기 혈압이 높게 측정되었다. **따라서 고혈압인 경우 관상동맥질환이 발병되기 쉽다고 판단할 수 있다.**




```
boxplot(diaBP~TenYearCHD,data=fram,col='red1')
```

#BMI와 TenYearCHD

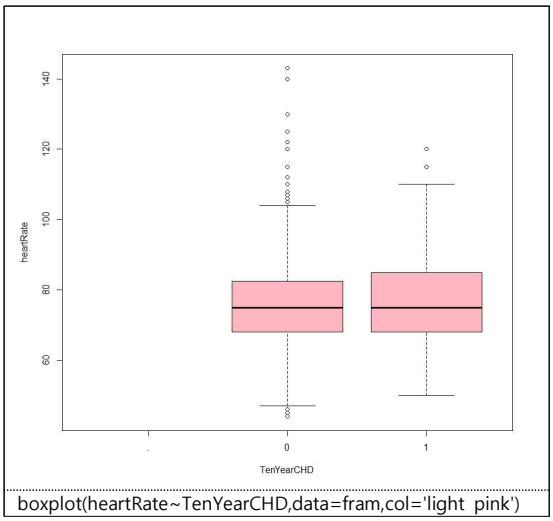
체질량 지수와 관상동맥질환 발병 여부에 대한 그래프는 다음과 같다. 중앙값이 비슷한 곳에 위치해 있으며 상자그림의 범위도 비슷하기 때문에 두 변수가 어떠한 관계를 가지고 있는지 파악하기 쉽지 않다.




```
boxplot(BMI~TenYearCHD,data=fram,col='medium  purple')
```

#heartRate와 TenYearCHD

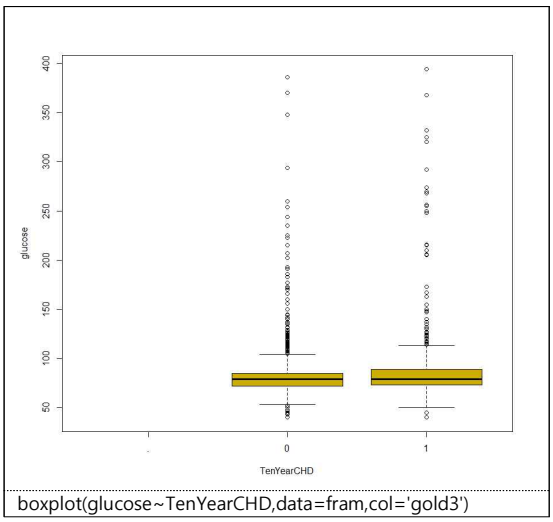
다음은 분당 심박수와 관상동맥질환 발병 여부에 관한 그래프이다. 상자그림이 유사하게 그려져 두 변수간 관계를 파악하기 쉽지않다.




```
boxplot(heartRate~TenYearCHD,data=fram,col='light  pink')
```

#glucose와 TenYearCHD

다음은 혈당 수치와 관상동맥질환 발병 여부에 관한 그래프이다. 마찬가지로 상자그림이 유사하게 그려져 두 변수간 관계를 파악하기 쉽지않다.




```
boxplot(glucose~TenYearCHD,data=fram,col='gold3')
```

#결론

결론적으로 TenYearCHD와 관련된 변수는 male, age, prevalentHyp, sysBP 그리고 diaBP가 있다. 여성보단 남성이, 나이가 많을수록, 우울증이 있을수록, 마지막으로 고혈압 일수록 십년 뒤 관상동맥질환이 발병되기 쉬울 것이다.