Assessment 1

ID: 10960269

Ans-1]

There could be few competing factors that could prevent an event from occurring, they are as follows:

1. In this prospective study we are interested in understanding the association between anti-hypertensive medication at exam time and CVD. If the medication has a positive impact on the participant then he/she is unlikely of developing CVD.
2. Censoring is a major factor that would prevent the event from occurring.

- A participant may not develop CVD before the study ends.
- A participant may be lost to follow-up during the study period and thereby will not get to know of his/her disease status.
- Death of the participant before developing CVD or not due to CVD would affect the end results.

3. For our analysis we need to make certain considerations as there are certain risk factors that could prevent the participant from developing CVD.

- Since CVD has long latency period and this study requires long period to analyze its participants so individuals having high cholesterol, belonging to older age-group, high blood pressure, diabetes are more likely to be lost during follow-up due to death or other reasons thereby preventing the event from happening and hence they require special considerations. One of the possible solutions is to avoid including very old participants in the study especially those having above mentioned diseases.

Ans-2]

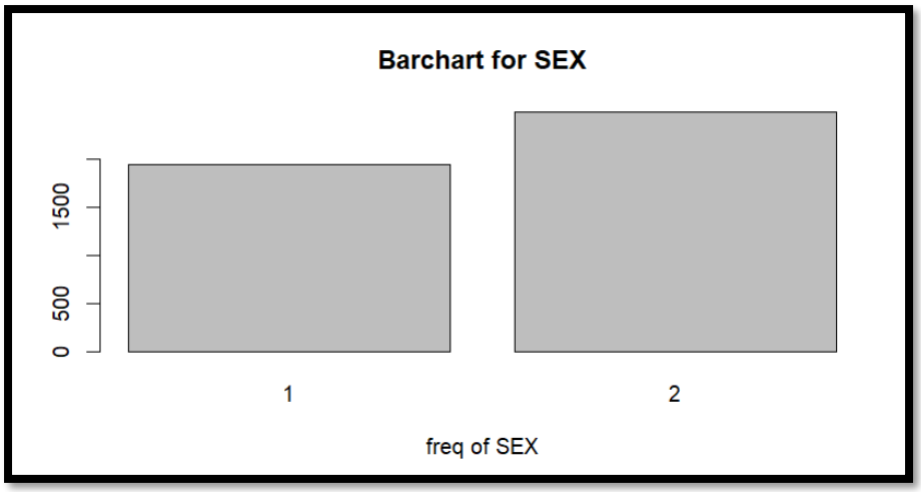
head(frmgham_data)

```
A tibble: 6 x 21
     ID    SEX TOTCHOL    AGE SYSBP DIABP CURSMOKE CIGPDAY   BMI DIABETES BPMEDS HEARTRTE GLUCOSE
  <int> <dbl>   <dbl> <dbl> <dbl> <dbl>    <dbl>   <dbl> <dbl>    <dbl>  <dbl>    <dbl>   <dbl>
      1     1     195    39  106     70        0       0  27.0        0      0       80      77
      2     2     250    46  121     81        0       0  28.7        0      0       95      76
      3     1     245    48  128.    80        1      20  25.3        0      0       75      70
      4     2     225    61  150     95        1      30  28.6        0      0       65     103
      5     2     285    46  130     84        1      23  23.1        0      0       85      85
      6     2     228    43  180    110        0       0  30.3        0      0       77      99
... with 8 more variables: educ <dbl>, PREVCHD <dbl>, DEATH <dbl>, CVD <dbl>, TIMECVD <dbl>,
  TIMEDTH <dbl>, cr_time <dbl>, cr_status <chr>
```

summary(frmgham_data)

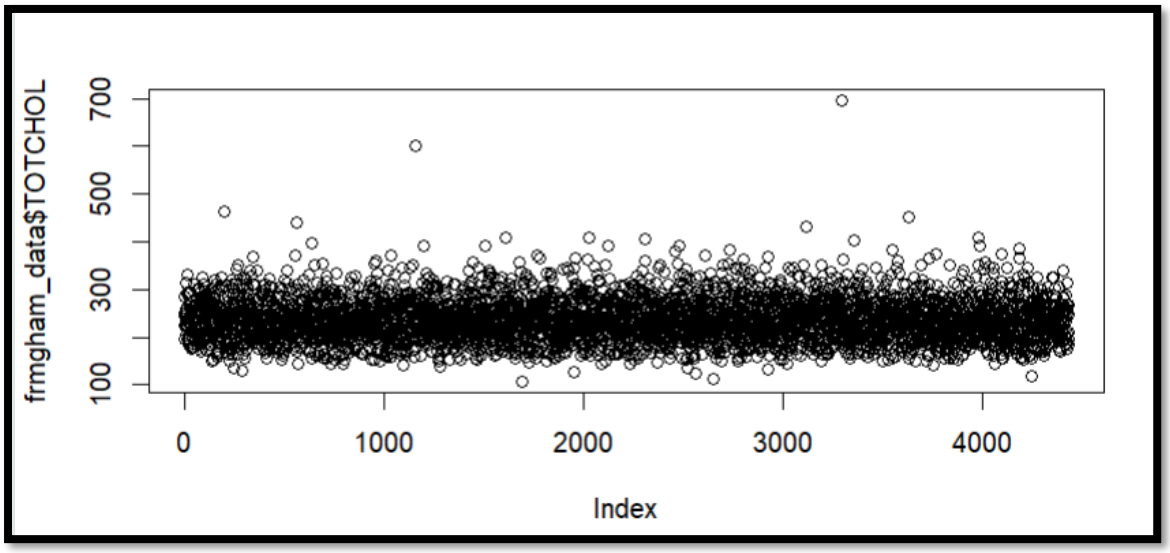
```
       ID              SEX            TOTCHOL          AGE            SYSBP           DIABP       
 Min.   :   1   Min.   :1.000   Min.   :107   Min.   :32.00   Min.   : 83.5   Min.   : 48.00  
 1st Qu.:1109   1st Qu.:1.000   1st Qu.:206   1st Qu.:42.00   1st Qu.:117.5   1st Qu.: 75.00  
 Median :2218   Median :2.000   Median :234   Median :49.00   Median :129.0   Median : 82.00  
 Mean   :2218   Mean   :1.562   Mean   :237   Mean   :49.93   Mean   :132.9   Mean   : 83.08  
 3rd Qu.:3326   3rd Qu.:2.000   3rd Qu.:264   3rd Qu.:57.00   3rd Qu.:144.0   3rd Qu.: 90.00  
 Max.   :4434   Max.   :2.000   Max.   :696   Max.   :70.00   Max.   :295.0   Max.   :142.50  
                                NA's   :52                                                    
    CURSMOKE         CIGPDAY           BMI           DIABETES          BPMEDS       
 Min.   :0.0000   Min.   : 0.000   Min.   :15.54   Min.   :0.00000   Min.   :0.00000  
 1st Qu.:0.0000   1st Qu.: 0.000   1st Qu.:23.09   1st Qu.:0.00000   1st Qu.:0.00000  
 Median :0.0000   Median : 0.000   Median :25.45   Median :0.00000   Median :0.00000  
 Mean   :0.4919   Mean   : 8.966   Mean   :25.85   Mean   :0.02729   Mean   :0.03293  
 3rd Qu.:1.0000   3rd Qu.:20.000   3rd Qu.:28.09   3rd Qu.:0.00000   3rd Qu.:0.00000  
 Max.   :1.0000   Max.   :70.000   Max.   :56.80   Max.   :1.00000   Max.   :1.00000  
                  NA's   :32       NA's   :19                        NA's   :61       
    HEARTRTE         GLUCOSE            educ          PREVCHD           DEATH        
 Min.   : 44.00   Min.   : 40.00   Min.   :1.000   Min.   :0.00000   Min.   :0.0000  
 1st Qu.: 68.00   1st Qu.: 72.00   1st Qu.:1.000   1st Qu.:0.00000   1st Qu.:0.0000  
 Median : 75.00   Median : 78.00   Median :2.000   Median :0.00000   Median :0.0000  
 Mean   : 75.89   Mean   : 82.19   Mean   :1.976   Mean   :0.04375   Mean   :0.3496  
 3rd Qu.: 83.00   3rd Qu.: 87.00   3rd Qu.:3.000   3rd Qu.:0.00000   3rd Qu.:1.0000  
 Max.   :143.00   Max.   :394.00   Max.   :4.000   Max.   :1.00000   Max.   :1.0000  
 NA's   :1        NA's   :397      NA's   :113                                       
      CVD            TIMECVD        TIMEDTH         cr_time      cr_status        
 Min.   :0.0000   Min.   :   0   Min.   :  26   Min.   :   0   Length:4434       
 1st Qu.:0.0000   1st Qu.:5145   1st Qu.:6974   1st Qu.:5134   Class :character  
 Median :0.0000   Median :8766   Median :8766   Median :8766   Mode  :character  
 Mean   :0.2609   Mean   :6818   Mean   :7506   Mean   :6814                     
 3rd Qu.:1.0000   3rd Qu.:8766   3rd Qu.:8766   3rd Qu.:8766                     
 Max.   :1.0000   Max.   :8766   Max.   :8766   Max.   :8766                     
```

barplot(table(frmgham_data$SEX) , xlab = 'freq of SEX', main = 'Barchart for SEX')



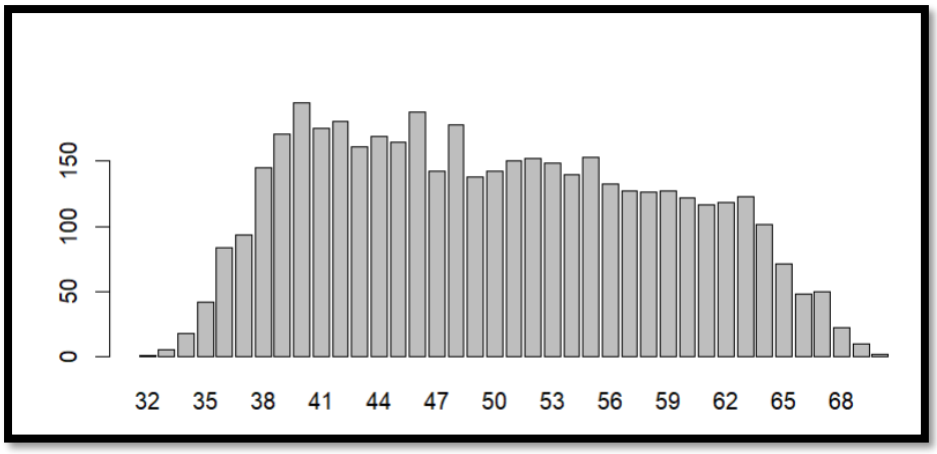
SEX is a categorical variable. The data in this column is discrete as shown by the barchart and hence will follow a discrete distribution. Women participants are more in this dataset.

```
plot(frmgham_data$TOTCHOL)
```



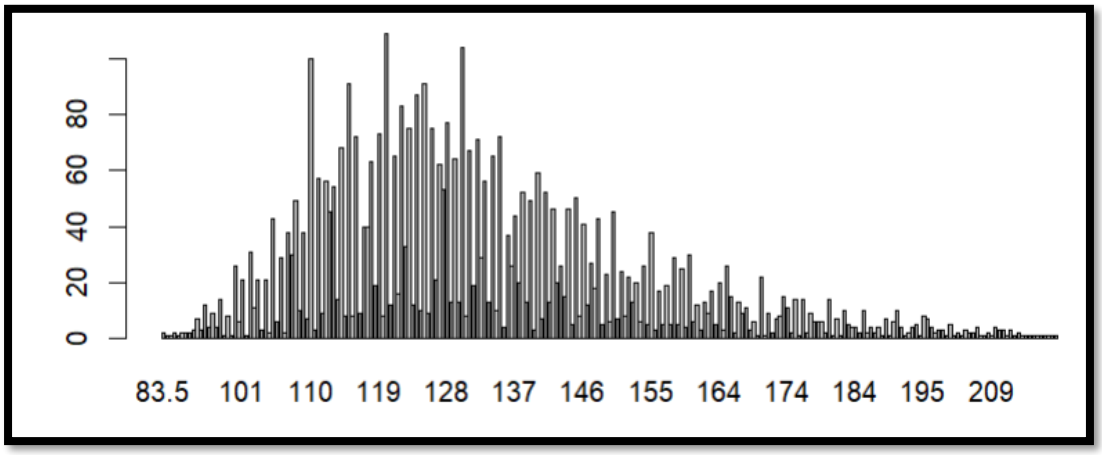
Cholesterol cannot be counted hence it is a continuous variable. The mean and median values differ slightly which indicates there is skewness in the data. Majority of the cholesterol values are below 300mg/dL.

```
barplot(table(frmgham_data$AGE))
```



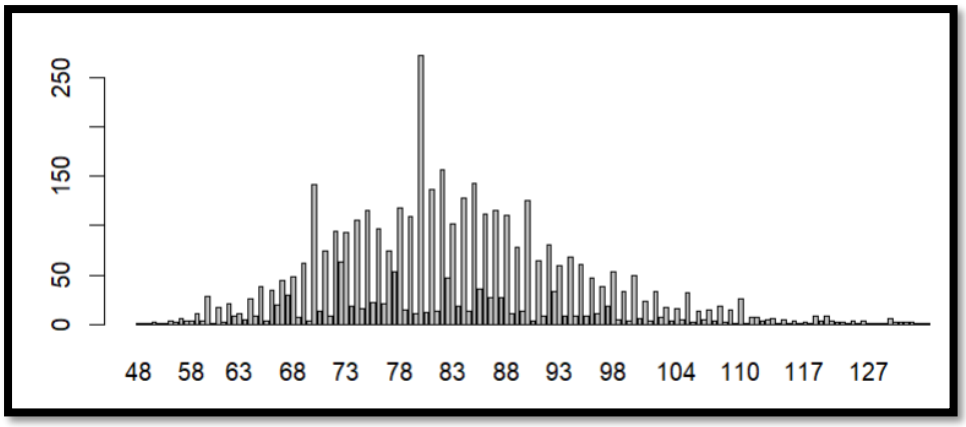
AGE is a continuous variable. The mean and median values differ very slightly which indicates there is very little skewness in the data.

```
barplot(table(frmgham_data$SYSBP))
```



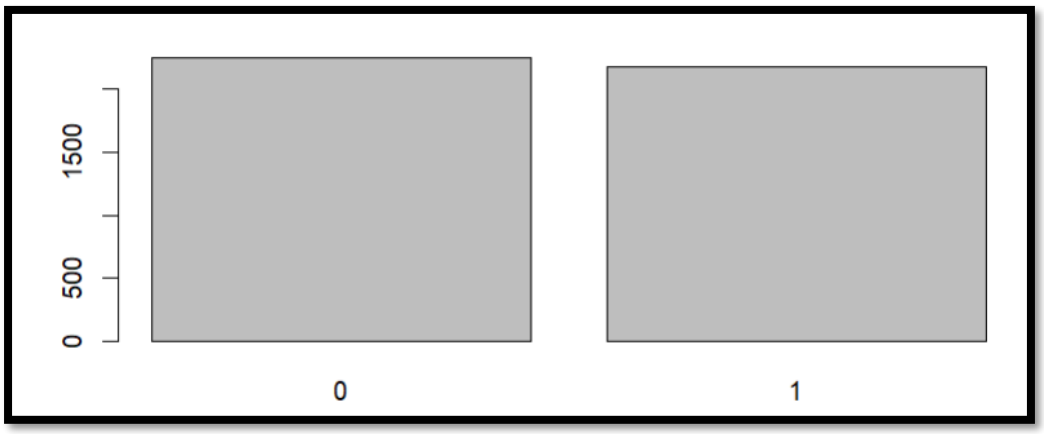
SYSBP is a continuous variable. The mean and median values differ slightly which indicates there is skewness in the data. Majority of the participants have systolic blood pressure varying between 107-152 mmHg.

```
barplot(table(frmgham_data$DIABP))
```



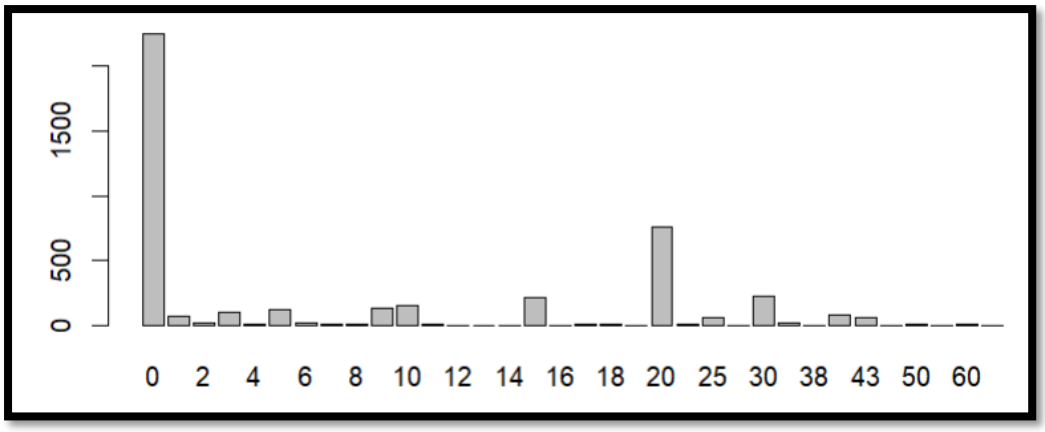
DIABP is a continuous variable. The mean and median values differ slightly which indicates there is skewness in the data. Majority of the participants have diastolic blood pressure varying between 68-93 mmHg.

```
barplot(table(frmgham_data$CURSMOKE))
```



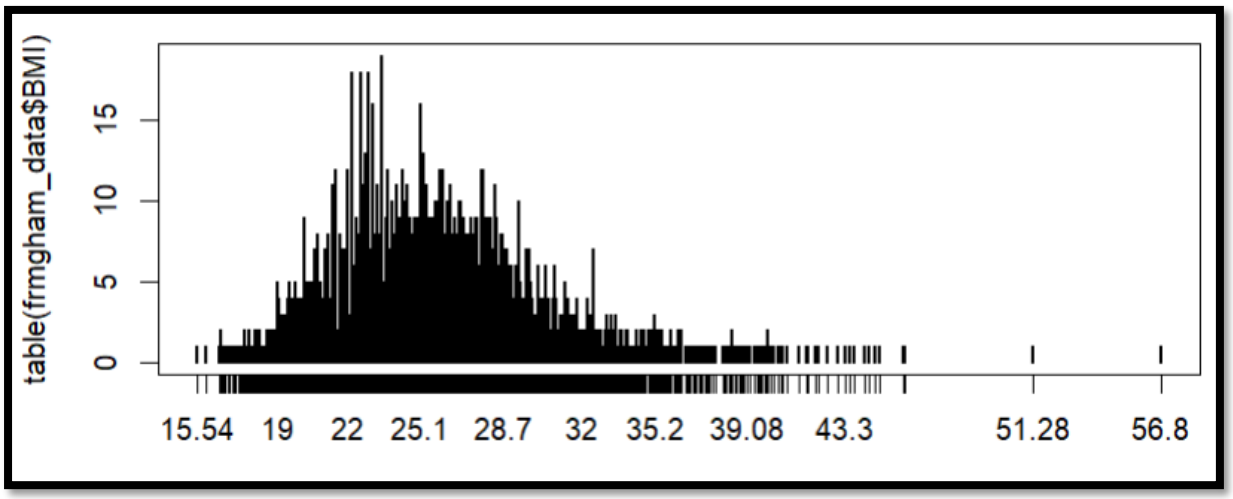
CURSMOKE is a categorical variable. The data in this column is discrete as shown by the barchart and hence will follow a discrete distribution. Frequency of non-current smokers are more.

```
barplot(table(frmgham_data$CIGPDAY))
```



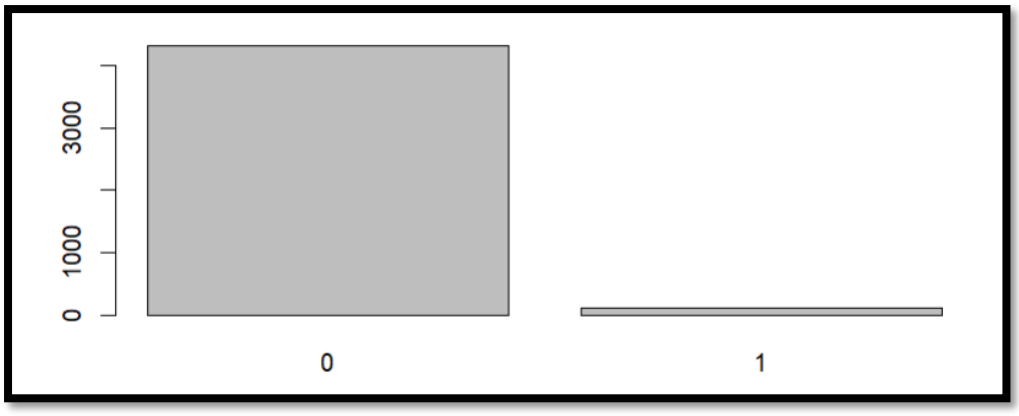
CIGPDAY is a categorical variable. The data in this column is discrete as shown by the barchart and hence will follow a discrete distribution. Frequency of non-current smokers are more.

```
plot(table(frmgham_data$BMI))
```



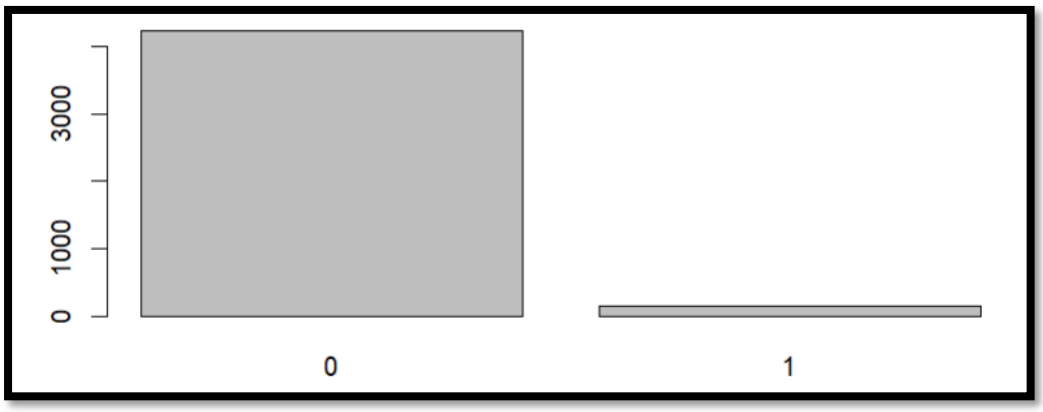
BMI is a continuous variable. The mean and median values differ very slightly which indicates there is very little skewness in the data. Majority of the participants have BMI varying between 19-32 weight in kilograms/height meters squared.

```
barplot(table(frmgham_data$DIABETES))
```



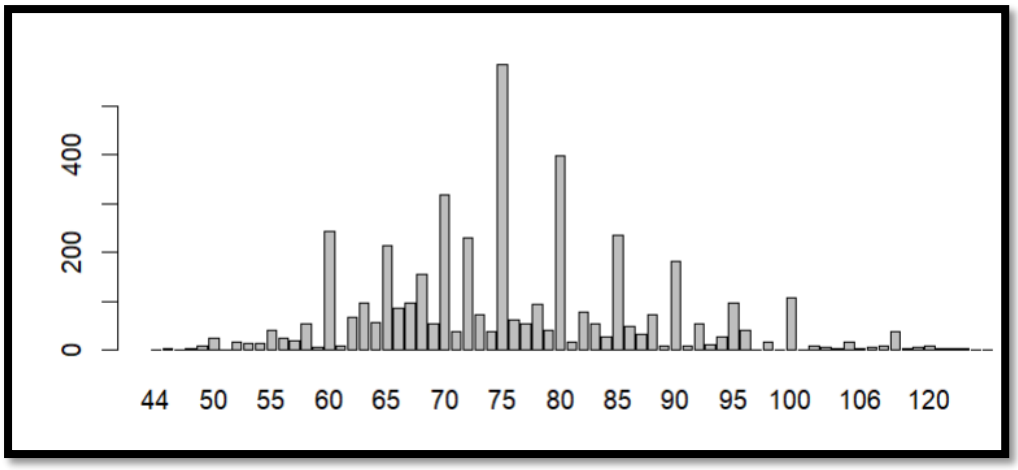
Diabetes is a categorical variable. The data in this column is discrete as shown by the barchart and hence will follow a discrete distribution. Frequency of non-diabetic participants are more.

```
barplot(table(frmgham_data$BPMEDS))
```



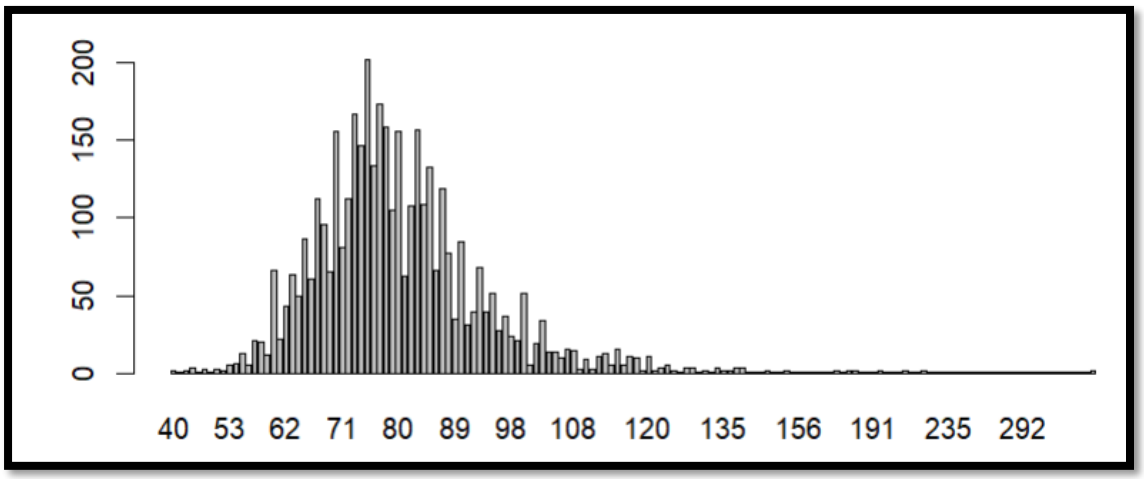
BPMEDS is a categorical variable. The data in this column is discrete as shown by the barchart and hence will follow a discrete distribution. Frequency of participants not using Anti-hypertensive medication at exam are more.

```
barplot(table(frmgham_data$HEARTRTE))
```



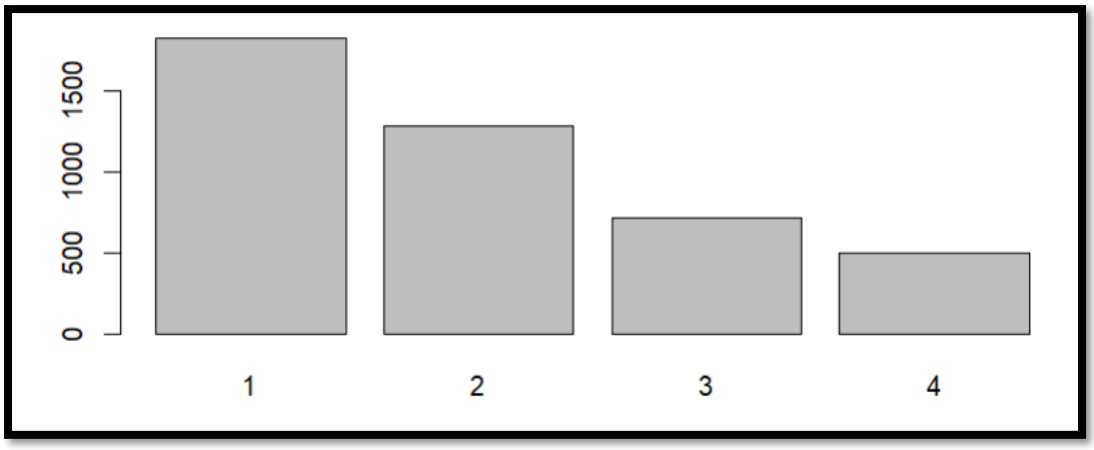
HEARTRTE is a continuous variable. The mean and median values differ very slightly which indicates there is very little skewness in the data. Heartrate of 75beats/min is significant.

```
barplot(table(frmgham_data$GLUCOSE))
```



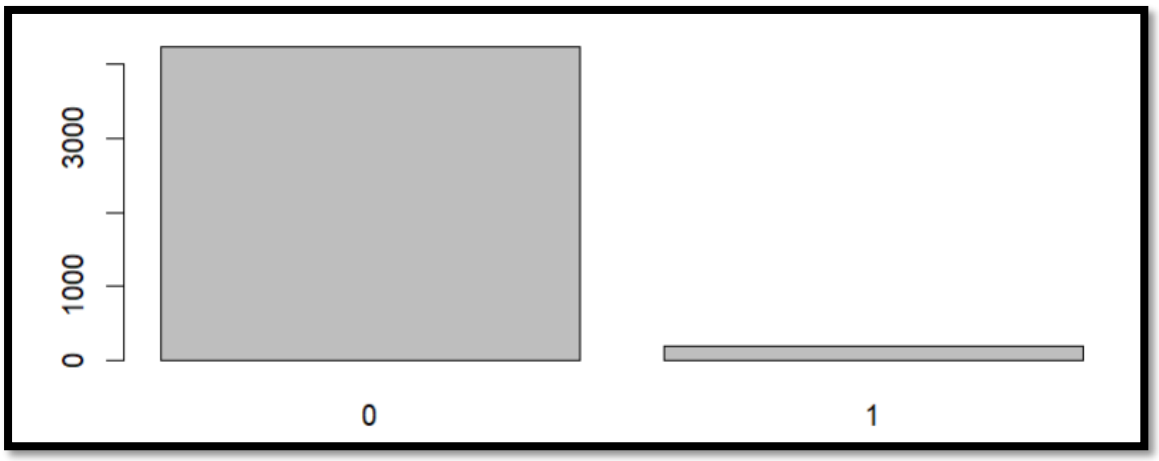
GLUCOSE is a continuous variable. The mean and median values differ slightly which indicates there is skewness in the data. Majority of the participants have glucose level varying between 62-90 mg/dL.

```
barplot(table(frmgham_data$educ))
```



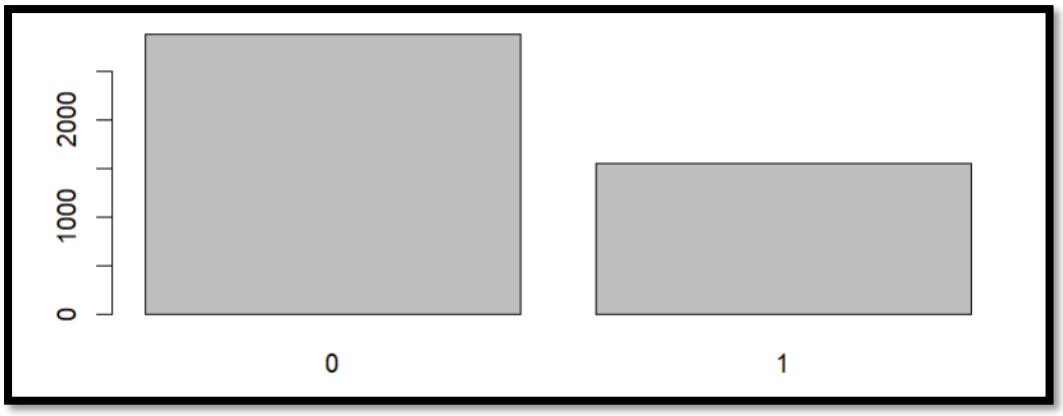
educ is a categorical variable. The data in this column is discrete as shown by the barchart and hence will follow a discrete distribution. Frequency of participants having 0-11 years of education are more.

```
barplot(table(frmgham_data$PREVCHD))
```



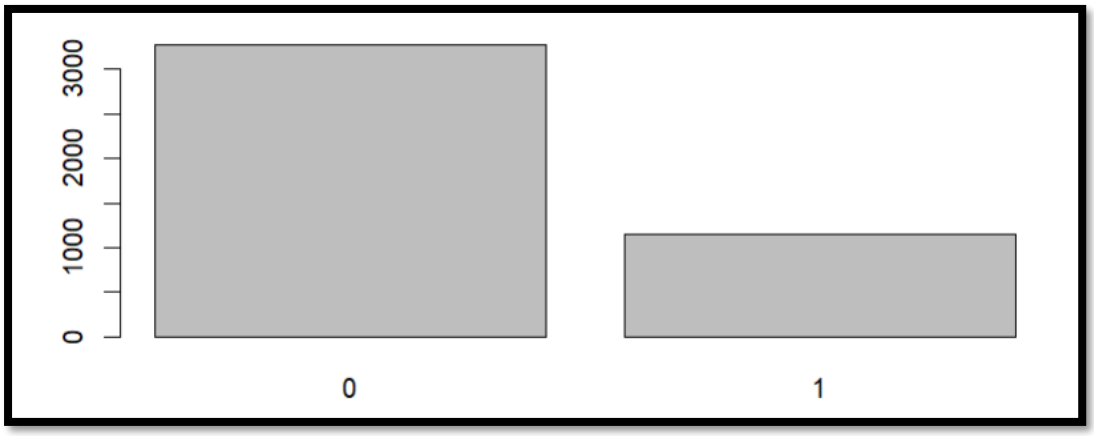
PREVCHD is a categorical variable. The data in this column is discrete as shown by the barchart and hence will follow a discrete distribution. Frequency of participants free of disease are more.

```
barplot(table(frmgham_data$DEATH))
```



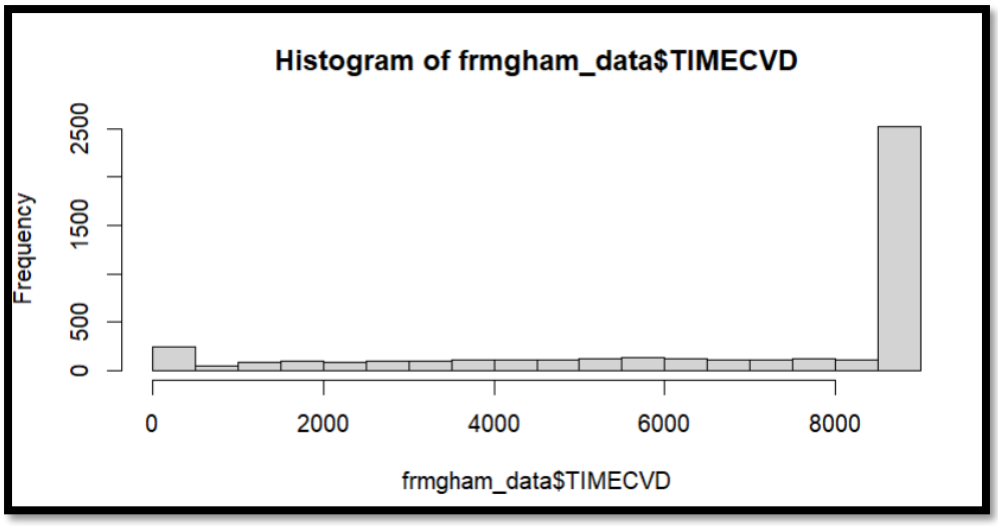
DEATH is a categorical variable. The data in this column is discrete as shown by the barchart and hence will follow a discrete distribution. Frequency of participants who did not die are more.

```
barplot(table(frmgham_data$CVD)
```



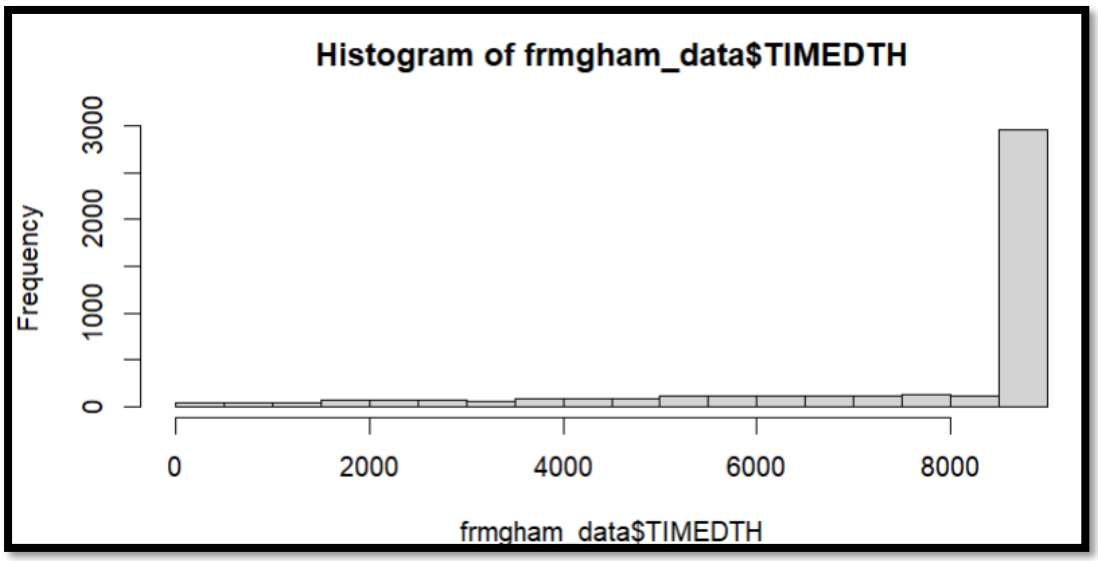
CVD is a categorical variable. The data in this column is discrete as shown by the barchart and hence will follow a discrete distribution. Frequency of participants who did not get CVD are more.

```
hist(frmgham_data$TIMECVD)
```



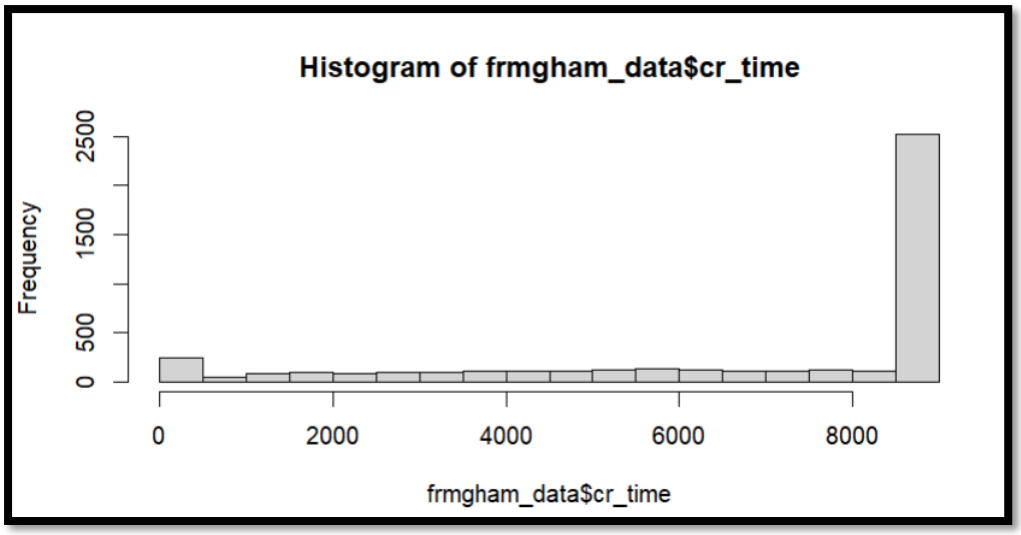
TIMECVD is a continuous variable. There is lot of skewness in the data as the mean and the median are different. The graph indicates that for majority of participants the first event of CVD during follow-up is noted beyond 8000 days.

```
hist(frmgham_data$TIMEDTH)
```



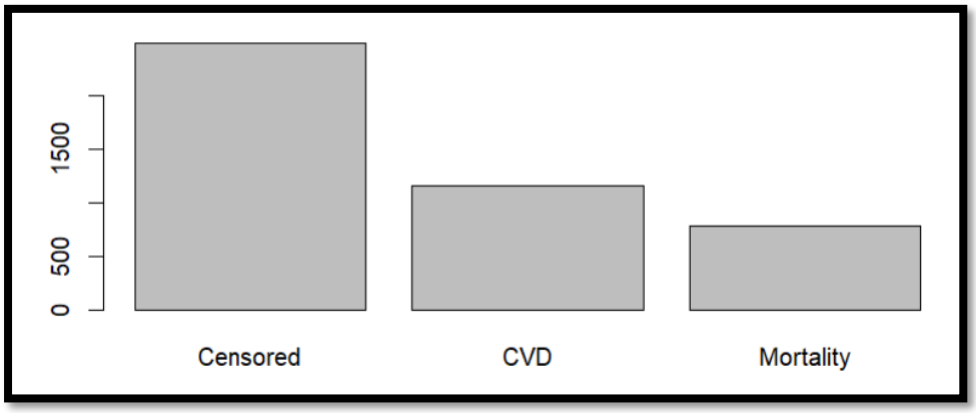
TIMEDTH is a continuous variable. There is lot of skewness in the data as the mean and the median are different. The graph indicates that for majority of participants deaths during follow-up is noted beyond 8000 days.

```
hist(frmgham_data$cr_time)
```



cr_time is a continuous variable. There is lot of skewness in the data as the mean and the median are different.

```
barplot(table(frmgham_data$cr_status))
```



cr_status is a categorical variable. The data in this column is discrete as shown by the barchart and hence will follow a discrete distribution. Frequency of participants who are censored are more.

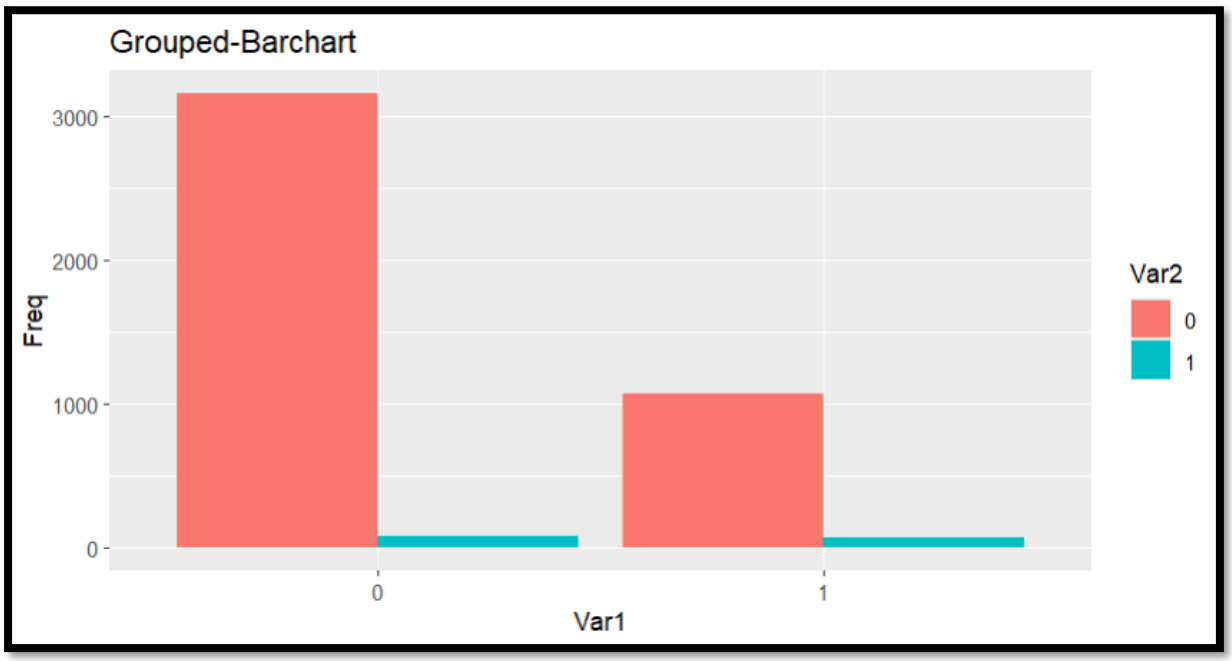
Since we are interested in understanding the association between variable BPMEDS and CVD, we perform exploratory data analysis on these variables.

```
x= table(frmgham_data$CVD,frmgham_data$BPMEDS)
data.frame(x)
```

```
data.frame(x)
Var1 Var2 Freq
   0    0 3163
   1    0 1066
   0    1   74
   1    1   70
```

```
ggplot(data.frame(x), aes(fill=Var2, x=Var1, y=Freq)) +
geom_bar(position="dodge", stat="identity")+ labs(title='Grouped-Barchart')
# used position = 'dodge' to get grouped barchart
```

Var1 = CVD ; Var2 = BPMEDS



Most of the participants who got CVD as well as who did not get CVD are those who did not use the medication.

Ans-3]

Information about the variables:

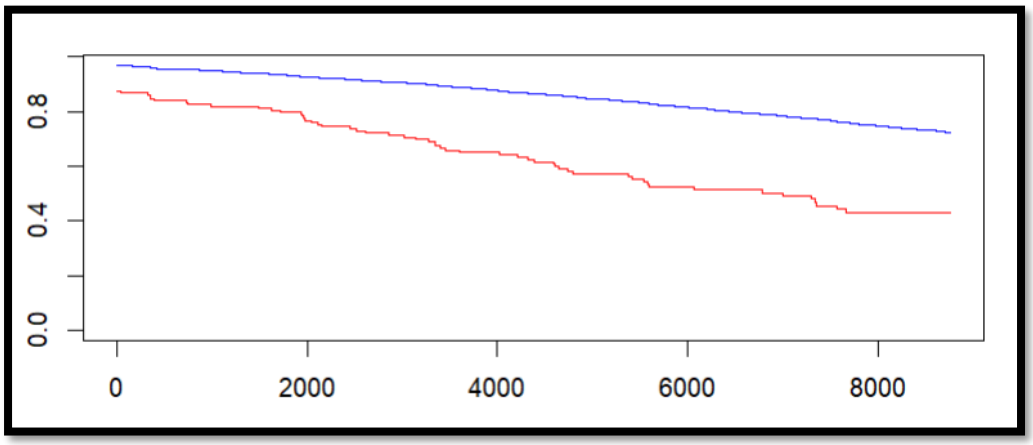
```
survfit(Surv(TIMECVD,CVD)~as.factor(BPMEDS),data =frmgham_data)
```

```
   61 observations deleted due to missingness
                        n events median 0.95LCL 0.95UCL
as.factor(BPMEDS)=0 4229   1066     NA      NA      NA
as.factor(BPMEDS)=1  144     70   7002    4800      NA
```

```
plot(survfit(Surv(TIMECVD,CVD)~as.factor(BPMEDS),data =frmgham_data),col = c('blue','red'))
```

Assumption: Non-informative censoring

Plotting Kaplan-Meier curves. The blue line indicates the survival function for 'BPMEDS = 0 ' group and red indicates the survival function for 'BPMEDS = 1 ' group.



The participants who do not use anti-hypertensive medication have better survival than those who take anti-hypertensive medication.

We perform log rank test to test the hypothesis that survival differs between two groups.

The hypothesis:

H0: no difference in survival between BPMEDS groups

H1: survival between BPMEDS groups differ

```
survdiff(Surv(TIMECVD,CVD)~as.factor(BPMEDS),data =frmgham_data)
```

```
Call:
survdiff(formula = Surv(TIMECVD, CVD) ~ as.factor(BPMEDS), data = frmgham_data)

n=4373, 61 observations deleted due to missingness.

                        N Observed Expected (O-E)^2/E (O-E)^2/V
as.factor(BPMEDS)=0 4229     1066   1111.1      1.83      84.1
as.factor(BPMEDS)=1  144       70     24.9     81.49      84.1

 Chisq= 84.1  on 1 degrees of freedom, p= <2e-16
>
```

p-value is less than 0.05 and hence we reject the null hypothesis and conclude that there is sufficient difference in survival distributions between both the groups of BPMEDS.

```
mcox = coxph(Surv(TIMECVD,CVD)~as.factor(BPMEDS),data =frmgham_data)
```

Assumption: Hazard function for all participants has same shape.

For performing non-competing risk analysis we fit Cox proportional hazard model using outcome variable TIMECVD and CVD to understand the association between use of anti-hypertensive medication at exam time with time-to-CVD.

```
summary(mcox)
```

```
Call:
coxph(formula = Surv(TIMECVD, CVD) ~ as.factor(BPMEDS), data = frmgham_data)

  n= 4373, number of events= 1136
   (61 observations deleted due to missingness)

                     coef exp(coef) se(coef)     z Pr(>|z|)
as.factor(BPMEDS)1 1.0871    2.9657   0.1237 8.787   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                   exp(coef) exp(-coef) lower .95 upper .95
as.factor(BPMEDS)1     2.966     0.3372     2.327     3.779

Concordance= 0.522  (se = 0.004 )
Likelihood ratio test= 57.29  on 1 df,   p=4e-14
Wald test            = 77.22  on 1 df,   p=<2e-16
Score (logrank) test = 85.11  on 1 df,   p=<2e-16
```

Interpretation:

(BPMEDS)1 variable has exp(coef) also referred to as hazard ratio.

At a given instant in time the participant who currently uses anti-hypertensive medication at exam is 2.966 times as likely to get CVD as someone who do not take the medication. We 95% confident that the hazard ratio is between the interval 2.327 to 3.779.

Model Selection:

```
fullmodel = coxph(formula = Surv(TIMECVD, CVD) ~ BPMEDS + SEX +
TOTCHOL+AGE+SYSBP+ DIABP+ CURSMOKE+CIGPDAY+BMI+DIABETES+
HEARTRTE+ GLUCOSE+ as.factor(educ)+ PREVCHD, data =
frmgham_data_complete)                                    # Full Model
```

```
summary(fullmodel)
```

```
                      coef  exp(coef)   se(coef)       z Pr(>|z|)
BPMEDS            0.0673867  1.0697090  0.1430410   0.471 0.637569
SEX              -0.8115151  0.4441846  0.0727301 -11.158  < 2e-16 ***
TOTCHOL           0.0036497  1.0036564  0.0007261   5.027 4.99e-07 ***
AGE               0.0511998  1.0525332  0.0044212  11.580  < 2e-16 ***
SYSBP             0.0138371  1.0139332  0.0022650   6.109 1.00e-09 ***
DIABP             0.0066146  1.0066365  0.0041589   1.590 0.111734
CURSMOKE          0.3367154  1.4003405  0.0988348   3.407 0.000657 ***
CIGPDAY           0.0040831  1.0040915  0.0039924   1.023 0.306439
BMI               0.0185115  1.0186839  0.0082415   2.246 0.024696 *
DIABETES          0.4824251  1.6199982  0.1764902   2.733 0.006268 **
HEARTRTE         -0.0083968  0.9916383  0.0027533  -3.050 0.002291 **
GLUCOSE           0.0041611  1.0041698  0.0011997   3.468 0.000524 ***
as.factor(educ)2  0.1200553  1.1275592  0.0789175   1.521 0.128191
as.factor(educ)3 -0.1492042  0.8613932  0.1012686  -1.473 0.140656
as.factor(educ)4 -0.0482198  0.9529243  0.1066538  -0.452 0.651186
PREVCHD           1.7205953  5.5878539  0.1066190  16.138  < 2e-16 ***
---
```

From the output we understand that the variables TOTCHOL, AGE, SYSBP, SEX CURSMOKE, BMI, DIABETES, HEARTRTE, GLUCOSE and PREVCHD are significant.

```
reduced = step(fullmodel)                                 # Optimal model selection
```

```
Call:
coxph(formula = Surv(TIMECVD, CVD) ~ SEX + TOTCHOL + AGE + SYSBP +
    DIABP + CURSMOKE + BMI + DIABETES + HEARTRTE + GLUCOSE +
    as.factor(educ) + PREVCHD, data = frmgham_data_complete)

  n= 3826, number of events= 1011

                       coef  exp(coef)   se(coef)       z Pr(>|z|)
SEX               -0.8259206  0.4378318  0.0704199 -11.729  < 2e-16 ***
TOTCHOL            0.0037019  1.0037087  0.0007233   5.118 3.09e-07 ***
AGE                0.0507381  1.0520474  0.0043929  11.550  < 2e-16 ***
SYSBP              0.0140529  1.0141521  0.0022288   6.305 2.88e-10 ***
DIABP              0.0065973  1.0066192  0.0041567   1.587 0.112476
CURSMOKE           0.4077447  1.5034233  0.0685062   5.952 2.65e-09 ***
BMI                0.0185418  1.0187148  0.0082420   2.250 0.024470 *
DIABETES           0.4757267  1.6091832  0.1761880   2.700 0.006932 **
HEARTRTE          -0.0082486  0.9917853  0.0027401  -3.010 0.002610 **
GLUCOSE            0.0041552  1.0041638  0.0012032   3.453 0.000554 ***
as.factor(educ)2   0.1204499  1.1280042  0.0788849   1.527 0.126784
as.factor(educ)3  -0.1464894  0.8637349  0.1012163  -1.447 0.147815
as.factor(educ)4  -0.0507659  0.9505011  0.1066252  -0.476 0.633992
PREVCHD            1.7260438  5.6183826  0.1051436  16.416  < 2e-16 ***
```

Thus we get a simplified model and from this model we understand that BPMEDS variable is not important and thereby indicating that there is no association between use of anti-hypertensive medication at exam time with time-to-CVD.

Proportional hazards assumption:

cox.zph(reduced)

```
                  chisq df      p
SEX               1.459  1 0.2270
TOTCHOL           3.960  1 0.0466
AGE               1.666  1 0.1968
SYSBP             0.121  1 0.7279
DIABP             0.224  1 0.6361
CURSMOKE          0.138  1 0.7099
BMI               0.067  1 0.7958
DIABETES          8.047  1 0.0046
HEARTRTE          0.744  1 0.3883
GLUCOSE           2.060  1 0.1512
as.factor(educ)   6.635  3 0.0845
PREVCHD         189.416  1 <2e-16
GLOBAL          218.600 14 <2e-16
```

PREVCHD variable has p-value less than 0.05 so coefficients value will happen to change overtime for this variable which could lead to a worse fitting model.

```
strat =  coxph(formula = Surv(TIMECVD, CVD) ~ SEX + TOTCHOL + AGE +
SYSBP + DIABP + CURSMOKE + BMI + DIABETES + HEARTRTE + GLUCOSE +
as.factor(educ) + strata(PREVCHD), data = frmgham_data_complete)
```

Performing stratification of PREVCHD variable for the solution of proportional hazards assumption's solution.

```
summary(strat)
```

```
                      coef  exp(coef)   se(coef)       z Pr(>|z|)
SEX              -0.8111123  0.4443635  0.0701031 -11.570  < 2e-16 ***
TOTCHOL           0.0036218  1.0036283  0.0007232   5.008 5.50e-07 ***
AGE               0.0496534  1.0509068  0.0044005  11.284  < 2e-16 ***
SYSBP             0.0140241  1.0141229  0.0022656   6.190 6.02e-10 ***
DIABP             0.0059854  1.0060034  0.0041819   1.431 0.152356
CURSMOKE          0.3521291  1.4220921  0.0678430   5.190 2.10e-07 ***
BMI               0.0240603  1.0243521  0.0082109   2.930 0.003386 **
DIABETES          0.4713454  1.6021483  0.1771676   2.660 0.007804 **
HEARTRTE         -0.0087156  0.9913222  0.0027192  -3.205 0.001350 **
GLUCOSE           0.0040454  1.0040536  0.0012200   3.316 0.000913 ***
as.factor(educ)2  0.0773410  1.0804104  0.0785530   0.985 0.324835
as.factor(educ)3 -0.1709696  0.8428472  0.1008907  -1.695 0.090151 .
as.factor(educ)4 -0.0631291  0.9388223  0.1065410  -0.593 0.553494
```

Thus, from the output we see PREVCHD variable is no longer considered important or significant.

Conclusion:

In the Cox proportional hazard model using BPMEDS as covariate indicated that the group 'BPMEDS = 1 ' i.e the participants who currently uses anti-hypertensive medication are more likely to get CVD than those who do not use the medication, thereby indicating there is an association between the BPMEDS and time to CVD. However, in the reduced model BPMEDS variable has no importance thereby contradicting the association.

Implications for competing risks:

1. Competing risks can lead to biased risk estimates if left ignored.
2. Participants who experience the competing event may have different prognosis for the primary event, compared to subjects who have neither censored nor experienced the competing event.

3. For example: Subjects who die of non-CVD causes may have a different prognosis than those who are currently event-free.

Ans-4]

We consider only those predictor variables in our Cox proportional hazards model that are recorded before the outcome such that it would help in predicting the outcome.

```
fullmodel2 = coxph(formula = Surv(TIMEDTH) ~ BPMEDS + SEX +
TOTCHOL+AGE+SYSBP+ DIABP+ CURSMOKE+CIGPDAY+BMI+DIABETES+
HEARTRTE+ GLUCOSE+ as.factor(educ)+ PREVCHD, data =
frmgham_data_complete)
```

```
summary(fullmodel2)
```

```
                    coef  exp(coef)   se(coef)      z Pr(>|z|)
BPMEDS           0.1992345  1.2204682  0.0952006  2.093  0.03637 *
SEX             -0.2420143  0.7850450  0.0363770 -6.653 2.87e-11 ***
TOTCHOL         -0.0007141  0.9992861  0.0004043 -1.766  0.07735 .
AGE              0.0298336  1.0302831  0.0023280 12.815  < 2e-16 ***
SYSBP            0.0074557  1.0074836  0.0013567  5.496 3.89e-08 ***
DIABP           -0.0025339  0.9974693  0.0023757 -1.067  0.28614
CURSMOKE         0.0701805  1.0727018  0.0524734  1.337  0.18108
CIGPDAY          0.0034238  1.0034297  0.0022706  1.508  0.13159
BMI             -0.0050713  0.9949416  0.0045645 -1.111  0.26656
DIABETES         0.4734324  1.6054954  0.1198870  3.949 7.85e-05 ***
HEARTRTE         0.0013240  1.0013249  0.0014083  0.940  0.34715
GLUCOSE          0.0028678  1.0028720  0.0008829  3.248  0.00116 **
as.factor(educ)2 0.0032157  1.0032208  0.0402305  0.080  0.93629
as.factor(educ)3 -0.0496176 0.9515933  0.0478107 -1.038  0.29937
as.factor(educ)4 -0.0950637 0.9093150  0.0543175 -1.750  0.08009 .
PREVCHD          0.4661602  1.5938623  0.0817770  5.700 1.20e-08 ***
---
```

From the output we see that we have seven significant variables.

Variable selection:

```
reduced2 = step(fullmodel2)          #performing backward model selection
summary(reduced2)
```

```
Call:
coxph(formula = Surv(TIMEDTH) ~ BPMEDS + SEX + TOTCHOL + AGE +
    SYSBP + CURSMOKE + CIGPDAY + DIABETES + GLUCOSE + PREVCHD,
    data = frmgham_data_complete)

  n= 3826, number of events= 3826

               coef  exp(coef)   se(coef)      z Pr(>|z|)
BPMEDS    0.1925985  1.2123960  0.0948005  2.032 0.042192 *
SEX      -0.2219472  0.8009577  0.0350864 -6.326 2.52e-10 ***
TOTCHOL  -0.0007627  0.9992376  0.0004018 -1.898 0.057652 .
AGE       0.0302526  1.0307149  0.0022380 13.518  < 2e-16 ***
SYSBP     0.0063345  1.0063546  0.0008609  7.358 1.87e-13 ***
CURSMOKE  0.0800107  1.0832987  0.0519606  1.540 0.123601
CIGPDAY   0.0036224  1.0036290  0.0022573  1.605 0.108540
DIABETES  0.4654558  1.5927400  0.1200601  3.877 0.000106 ***
GLUCOSE   0.0029153  1.0029196  0.0008805  3.311 0.000930 ***
PREVCHD   0.4678982  1.5966349  0.0812331  5.760 8.41e-09 ***
---
```

Thus, we get the optimal model with most significant variables.

Report:

Title:

Development of model for prediction of death due to various covariates to estimate future risk of cardiovascular disease: prospective cohort study.

Abstract:

Design: Prospective Cohort Study.

Setting: The data in the attached file is a subset of the Framingham study's data and contains laboratory, clinic, questionnaire, and adjudicated event data for 4,434 individuals.

Participants: From around 1956 to 1968, participant clinic data was gathered throughout three evaluation sessions, each about six years apart. Each participant was tracked for a total of 24 years to see how the following events played out: Angina Pectoris, Myocardial Infarction, Atherothrombotic Infarction, Cerebral Hemorrhage (Stroke), or death are all possible outcomes.

Methods: Used cox proportional hazards model for prediction and performed stepAIC() backward model selection for variable selection.

Outcome: 'TIMEDTH' - Number of days from Baseline exam to death if occurring during follow-up or Number of days from Baseline to censor date.

Predictors: By backward model selection we selected our variables as- BPMEDS, SEX, TOTCHOL, AGE, SYSBP, CURSMOKE, CIGPDAY, DIABETES, GLUCOSE and PREVCHD.

Missing data: Missing data was handled using the command

```
frmgham_data_complete <- frmgham_data %>%
 filter(complete.cases(.))
```

Ans-5]

1. Generally area under the curve of 0.7 or higher is very good for a predictive model.
2. So, the CPM would efficiently separate participants who develop an outcome from those who will not.
3. Since the performance of this model is on the same data that was used to develop it, these high performances could be due to in-sample optimization or overfitting of the dataset.
4. The overfitting can be avoided by performing internal validation by using ways like Split-Sample or Cross-validation.
5. By using these ways we make sure to evaluate the performance of the model on the unknown data(Validation sample of the dataset) thereby minimizing the chances of overfitting.
6. In Split-Sample method, we divide the dataset such that we develop the CPM using around 75% of the data and the remaining 25% is used for validating the CPM thereby minimizing overfitting.
7. In Cross-validation method, we divide the dataset into k-folds and test/validate the CPM using every fold and lastly take the average of

across all the folds. This is a very efficient way of performing internal validation to minimize overfitting.

8. Thus internal validation can be used to test the performance of CPM in data that are similar to the development dataset.
9. Hence, such results are only informative if we perform internal validation of the model.

Ans-6]

The steps to address these issues are:

1. We can do model updating upon the existing prediction model by adding more covariates.
2. To suite the external population we can change the part of the formula we are using in the model.
3. We can tweak the intercepts in our model or other parameters and test the model till we get an optimal performance.
4. We can discard the model and build a new one.
5. We can report or publish the model focusing on its performance parameters and highlighting its limitations.