R Notebook

The following is your first chunk to start with. Remember, you can add chunks using the menu above (Insert -> R) or using the keyboard shortcut Ctrl+Alt+I. A good practice is to use different code chunks to answer different questions. You can delete this comment if you like.

Other useful keyboard shortcuts include Alt- for the assignment operator, and Ctrl+Shift+M for the pipe operator. You can delete these reminders if you don't want them in your report.

```r
setwd("C:/") #Don't forget to set your working directory before you start!

library("tidyverse")
```

```
## -- Attaching packages -----------------------------------------------
-----------------------------------------------------------------------
- tidyverse 1.3.0 --

## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## -- Conflicts ----------------------------------------------------------
-----------------------------------------------------------------------
tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library("tidymodels")
```

```
## Registered S3 method overwritten by 'xts':
##   method     from
##   as.zoo.xts zoo

## -- Attaching packages -----------------------------------------------
-----------------------------------------------------------------------
tidymodels 0.0.3 --

## v broom     0.5.3     v recipes   0.1.9
## v dials     0.0.4     v rsample   0.0.5
## v infer     0.5.1     v yardstick 0.0.4
## v parsnip   0.0.5

## -- Conflicts ----------------------------------------------------------
----------------------------------------------------------------------
tidymodels_conflicts() --
## x scales::discard()   masks purrr::discard()
## x dplyr::filter()     masks stats::filter()
## x recipes::fixed()    masks stringr::fixed()
## x dplyr::lag()        masks stats::lag()
## x dials::margin()     masks ggplot2::margin()
## x yardstick::spec()   masks readr::spec()
## x recipes::step()     masks stats::step()
## x recipes::yj_trans() masks scales::yj_trans()
```

```r
library("plotly")
```

```
##
## Attaching package: 'plotly'
```

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

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
## The following object is masked from 'package:graphics':
##
##     layout
```

```r
library("skimr")
library("caret")
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

```
## The following objects are masked from 'package:yardstick':
##
##     precision, recall
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```r
dff= read_csv("lab3FraminghamHeart.csv")
```

```
## Parsed with column specification:
## cols(
##   gender = col_double(),
##   age = col_double(),
##   education = col_double(),
##   currentSmoker = col_double(),
##   cigsPerDay = col_double(),
##   BPMeds = col_double(),
##   prevalentStroke = col_double(),
##   prevalentHyp = col_double(),
##   diabetes = col_double(),
##   totChol = col_double(),
##   sysBP = col_double(),
##   diaBP = col_double(),
##   BMI = col_double(),
##   heartRate = col_double(),
##   glucose = col_double(),
##   TenYearCHD = col_double()
## )
```

```
colsToFactor <- c('gender', 'education', 'currentSmoker', 'BPMeds',
'prevalentStroke', 'prevalentHyp', 'diabetes')

dff <- dff %>%
  mutate_at(colsToFactor, ~factor(.))
```

WHAT DO YOU THINK MUTATE_AT DOES?

```
str(dff)

## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 3658 obs. of  16
variables:
##  $ gender         : Factor w/ 2 levels "0","1": 2 1 2 1 1 1 1 1 2 2 ...
##  $ age            : num  39 46 48 61 46 43 63 45 52 43 ...
##  $ education      : Factor w/ 4 levels "1","2","3","4": 4 2 1 3 3 2 1 2 1
1 ...
##  $ currentSmoker  : Factor w/ 2 levels "0","1": 1 1 2 2 2 1 1 2 1 2 ...
##  $ cigsPerDay     : num  0 0 20 30 23 0 0 20 0 30 ...
##  $ BPMeds         : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ prevalentStroke: Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ prevalentHyp   : Factor w/ 2 levels "0","1": 1 1 1 2 1 2 1 1 2 2 ...
##  $ diabetes       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ totChol        : num  195 250 245 225 285 228 205 313 260 225 ...
##  $ sysBP          : num  106 121 128 150 130 ...
##  $ diaBP          : num  70 81 80 95 84 110 71 71 89 107 ...
##  $ BMI            : num  27 28.7 25.3 28.6 23.1 ...
##  $ heartRate      : num  80 95 75 65 85 77 60 79 76 93 ...
##  $ glucose        : num  77 76 70 103 85 99 85 78 79 88 ...
##  $ TenYearCHD     : num  0 0 0 1 0 0 1 0 0 0 ...
```

Question 1

```
plotQ11 <- dff %>%
  ggplot(aes(x= TenYearCHD, y=sysBP, group= TenYearCHD))+
  geom_boxplot()
ggplotly(plotQ11)
```

The above boxplot is for sysBP

```
plotQ12 <- dff %>%
  ggplot(aes(x= TenYearCHD, y=diaBP, group= TenYearCHD)) +
  geom_boxplot()
ggplotly(plotQ12)
```

The above boxplot is for diaBP

```
plotQ13 <- dff %>%
  ggplot(aes(x= TenYearCHD, y=totChol, group= TenYearCHD)) +
  geom_boxplot()
ggplotly(plotQ13)
```

The above boxplot is for totChol

Question 2 part i

```
set.seed(123)
dffTrain <- dff %>%  sample_frac(0.7)
dffTest <- dplyr::setdiff(dff,dffTrain)
```

Question 2 part ii

Gender:

```
dffTrain %>% group_by(gender) %>%
  tally() %>%
  mutate(pct = 100*n/sum(n))
```

```
## # A tibble: 2 x 3
##   gender     n   pct
##   <fct>  <int> <dbl>
## 1 0       1419  55.4
## 2 1       1142  44.6
```

```
dffTest %>% group_by(gender) %>%
  tally() %>%
  mutate(pct = 100*n/sum(n))
```

```
## # A tibble: 2 x 3
##   gender     n   pct
##   <fct>  <int> <dbl>
## 1 0        616  56.2
## 2 1        481  43.8
```

Age:

```
dffTrain %>% group_by(ageGroup=cut_interval(age, length=10)) %>%
  tally() %>%
  mutate(pct = 100*n/sum(n))
```

```
## # A tibble: 4 x 3
##   ageGroup     n   pct
##   <fct>    <int> <dbl>
## 1 [30,40]    467  18.2
## 2 (40,50]    973  38.0
## 3 (50,60]    772  30.1
## 4 (60,70]    349  13.6
```

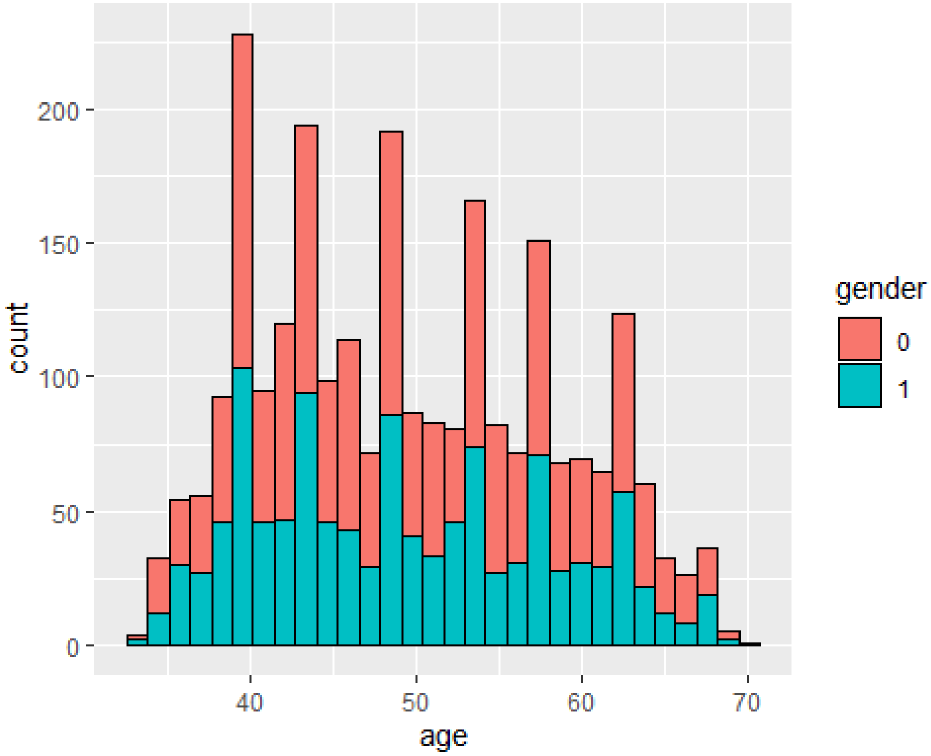
```
dffTest %>% group_by(ageGroup=cut_interval(age, length=10)) %>%
  tally() %>%
  mutate(pct = 100*n/sum(n))
```

```
## # A tibble: 4 x 3
##   ageGroup     n   pct
##   <fct>    <int> <dbl>
## 1 [30,40]    181  16.5
## 2 (40,50]    421  38.4
## 3 (50,60]    346  31.5
## 4 (60,70]    149  13.6
```

For Histogram:

```
plotQ2 <- dffTrain %>%
  ggplot(aes(x=age, fill=gender)) +
  geom_histogram(color='black')
plotQ2
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Question 3

```
fitLPM <- lm(TenYearCHD ~., data= dffTrain)
summary(fitLPM)
```

```
## 
## Call:
## lm(formula = TenYearCHD ~ ., data = dffTrain)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.69588 -0.18760 -0.09864 -0.00854  1.06563 
## 
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      -0.5193243  0.0939086  -5.530 3.53e-08 ***
## gender1           0.0402834  0.0149552   2.694  0.00711 ** 
## age               0.0073056  0.0009204   7.938 3.06e-15 ***
## education2       -0.0114841  0.0167200  -0.687  0.49224    
## education3       -0.0345910  0.0196551  -1.760  0.07854 .  
## education4       -0.0259428  0.0230652  -1.125  0.26080    
## currentSmoker1    0.0143681  0.0216179   0.665  0.50634    
## cigsPerDay        0.0018669  0.0009316   2.004  0.04519 *  
## BPMeds1           0.0184297  0.0434995   0.424  0.67184    
## prevalentStroke1  0.2099878  0.0983542   2.135  0.03285 *  
## prevalentHyp1     0.0448001  0.0208879   2.145  0.03206 *  
## diabetes1         0.0204464  0.0513727   0.398  0.69066    
## totChol           0.0002882  0.0001590   1.813  0.07000 .  
## sysBP             0.0023876  0.0005798   4.118 3.95e-05 ***
## diaBP            -0.0016597  0.0009716  -1.708  0.08770 .  
## BMI               0.0007242  0.0018265   0.397  0.69175    
## heartRate        -0.0013046  0.0005843  -2.233  0.02566 *  
## glucose           0.0011775  0.0003608   3.264  0.00111 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.3388 on 2543 degrees of freedom
## Multiple R-squared:  0.1077, Adjusted R-squared:  0.1017 
## F-statistic: 18.05 on 17 and 2543 DF,  p-value: < 2.2e-16
```

```
car::vif(fitLPM)
```

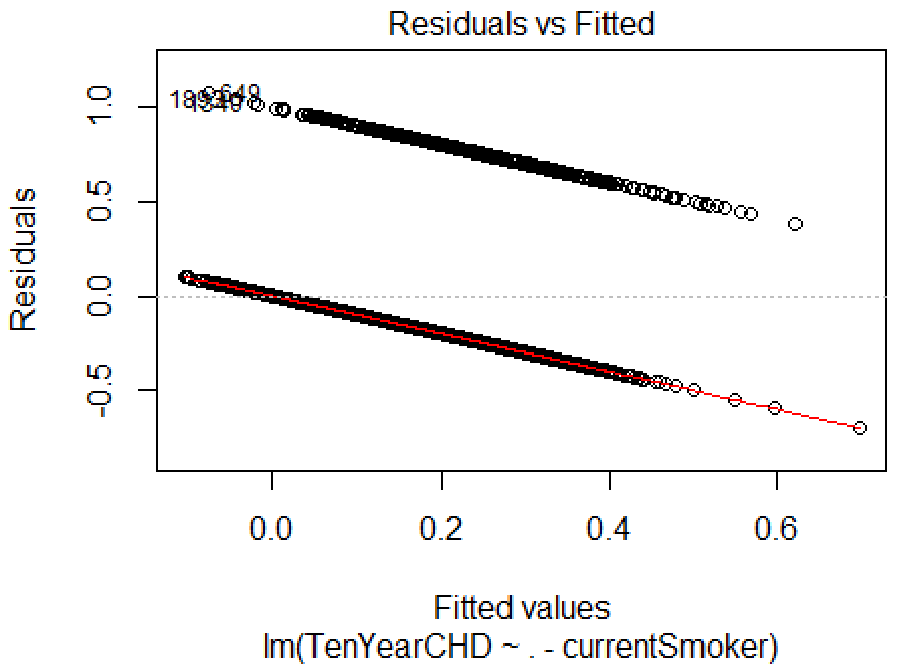
```
## Registered S3 methods overwritten by 'car':
##   method                          from
##   influence.merMod                lme4
##   cooks.distance.influence.merMod lme4
##   dfbeta.influence.merMod         lme4
##   dfbetas.influence.merMod        lme4
```

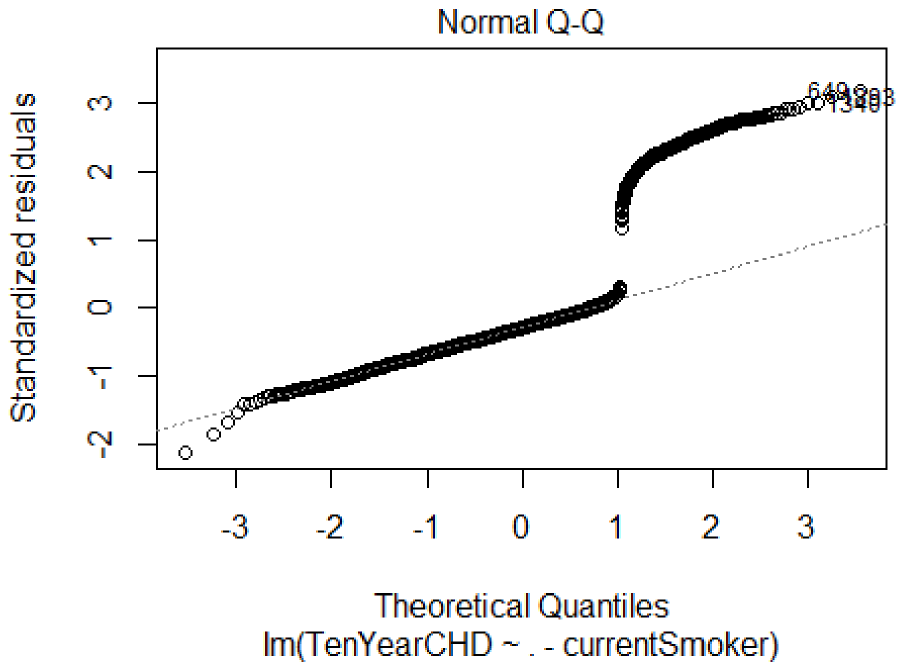
```
##                     GVIF Df GVIF^(1/(2*Df))
## gender          1.232950  1        1.110383
## age             1.398367  1        1.182526
## education       1.139817  3        1.022051
## currentSmoker   2.604754  1        1.613925
## cigsPerDay      2.762784  1        1.662163
## BPMeds          1.106826  1        1.052058
## prevalentStroke 1.006585  1        1.003287
## prevalentHyp    2.057398  1        1.434363
## diabetes        1.630615  1        1.276956
## totChol         1.106930  1        1.052107
## sysBP           3.777158  1        1.943491
## diaBP           2.997947  1        1.731458
## BMI             1.227604  1        1.107973
## heartRate       1.095878  1        1.046842
## glucose         1.645722  1        1.282857
```

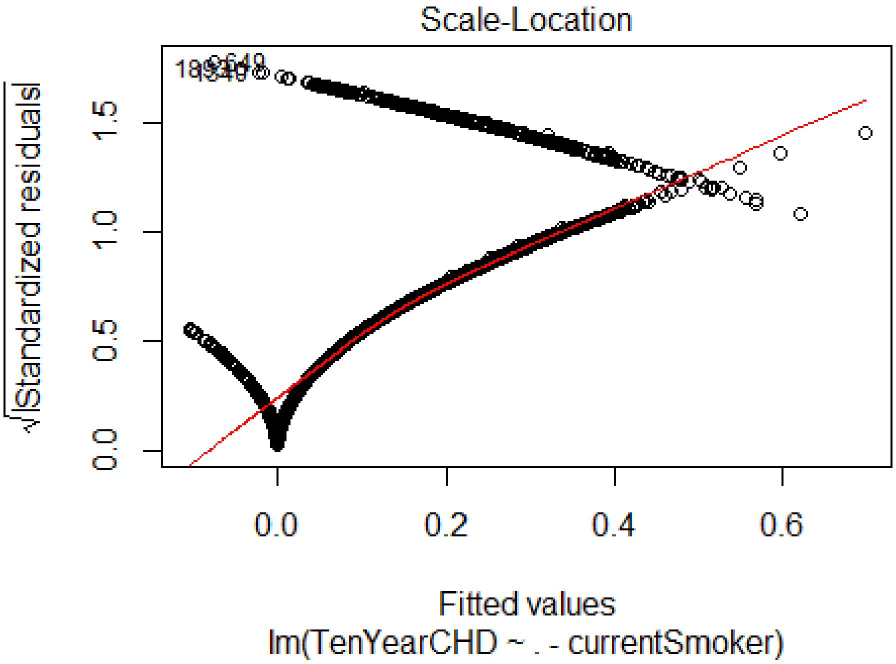
```
newModelfitLPM <- lm(TenYearCHD ~. -currentSmoker, data= dffTrain)
summary(newModelfitLPM)
```

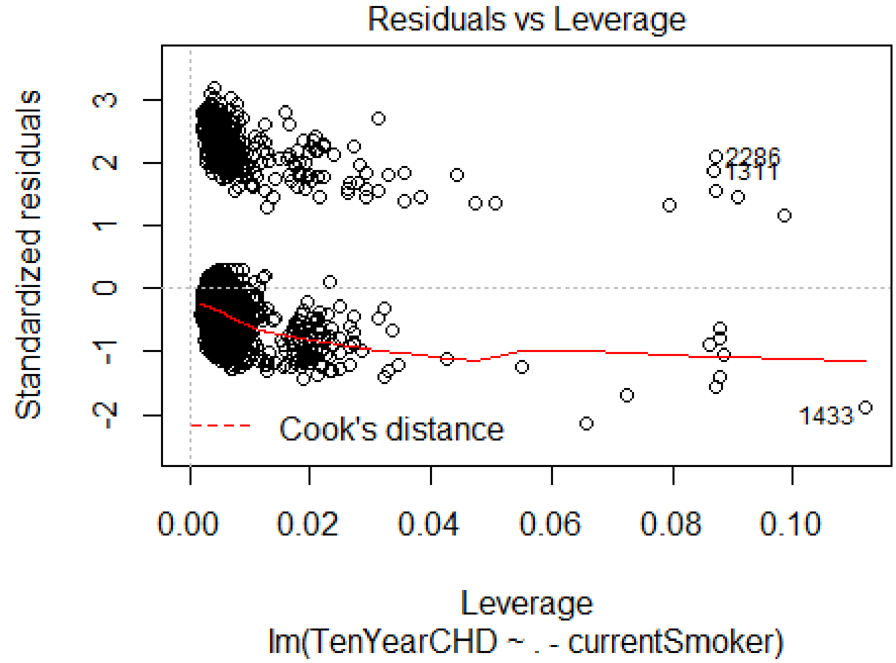
```
## 
## Call:
## lm(formula = TenYearCHD ~ . - currentSmoker, data = dffTrain)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.69721 -0.18848 -0.09967 -0.00937  1.07518 
## 
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      -0.5092583  0.0926691  -5.495 4.28e-08 ***
## gender1           0.0396262  0.0149208   2.656 0.007962 ** 
## age               0.0072591  0.0009176   7.911 3.78e-15 ***
## education2       -0.0113009  0.0167159  -0.676 0.499067    
## education3       -0.0346151  0.0196529  -1.761 0.078304 .  
## education4       -0.0260964  0.0230615  -1.132 0.257909    
## cigsPerDay        0.0023323  0.0006145   3.795 0.000151 ***
## BPMeds1           0.0185984  0.0434940   0.428 0.668972    
## prevalentStroke1  0.2097097  0.0983425   2.132 0.033066 *  
## prevalentHyp1     0.0448426  0.0208855   2.147 0.031882 *  
## diabetes1         0.0203925  0.0513670   0.397 0.691403    
## totChol           0.0002875  0.0001590   1.809 0.070633 .  
## sysBP             0.0023882  0.0005798   4.119 3.92e-05 ***
## diaBP            -0.0016833  0.0009708  -1.734 0.083051 .  
## BMI               0.0006191  0.0018194   0.340 0.733670    
## heartRate        -0.0013019  0.0005843  -2.228 0.025944 *  
## glucose           0.0011752  0.0003607   3.258 0.001138 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.3388 on 2544 degrees of freedom
## Multiple R-squared:  0.1075, Adjusted R-squared:  0.1019 
## F-statistic: 19.16 on 16 and 2544 DF,  p-value: < 2.2e-16
```

```
plot(newModelfitLPM)
```


Residuals vs Fitted
Residuals
1.0
0.5
0.0
-0.5
1893
649
1349
0.0
0.2
0.4
0.6
Fitted values
lm(TenYearCHD ~ . - currentSmoker)


Normal Q-Q
Standardized residuals
3
2
1
0
-1
-2
649
1349
1893
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(TenYearCHD ~ . - currentSmoker)





Question 4

```
resultsLPM <-
    lm( TenYearCHD ~. -currentSmoker, data= dffTrain ) %>%
    predict(., dffTest) %>%
    bind_cols(dffTest, predictedProb=.) %>%
    mutate(predictedClass = ifelse(predictedProb > 0.5, 1, 0))
resultsLPM

## # A tibble: 1,097 x 18
##    gender   age education currentSmoker cigsPerDay BPMeds prevalentStroke
##    <fct>  <dbl> <fct>     <fct>              <dbl> <fct>  <fct>
##  1 1         48 1         1                     20 0      0
##  2 0         43 2         0                      0 0      0
##  3 0         43 2         0                      0 0      0
##  4 0         41 3         0                      0 1      0
##  5 0         52 3         1                     20 0      0
##  6 0         61 3         0                      0 0      0
##  7 1         46 1         1                     20 0      0
##  8 0         63 2         1                     40 0      0
##  9 0         62 1         0                      0 0      0
## 10 1         49 1         1                      2 0      0
## # ... with 1,087 more rows, and 11 more variables: prevalentHyp <fct>,
## #   diabetes <fct>, totChol <dbl>, sysBP <dbl>, diaBP <dbl>, BMI <dbl>,
## #   heartRate <dbl>, glucose <dbl>, TenYearCHD <dbl>, predictedProb <dbl>,
## #   predictedClass <dbl>

dffTest %>% group_by(TenYearCHD) %>%
  tally() %>%
  mutate(pct = 100*n/sum(n))

## # A tibble: 2 x 3
##   TenYearCHD     n   pct
##        <dbl> <int> <dbl>
## 1          0   925  84.3
## 2          1   172  15.7

resultsLPM %>% group_by(predictedClass) %>%
  tally() %>%
  mutate(pct = 100*n/sum(n))

## # A tibble: 2 x 3
##   predictedClass     n    pct
##            <dbl> <int>  <dbl>
## 1              0  1087 99.1
## 2              1    10  0.912

dffTest <- dffTest %>%
  mutate(TenYearCHD = as.factor(TenYearCHD))
dffTrain <- dffTrain %>%
  mutate(TenYearCHD = as.factor(TenYearCHD))
```

Question 5

```
fitGLMQ5 <- glm(TenYearCHD ~. -currentSmoker, family = binomial(), data=
dffTrain)
summary(fitGLMQ5)
```

```
##
## Call:
## glm(formula = TenYearCHD ~ . - currentSmoker, family = binomial(),
##     data = dffTrain)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8022  -0.5882  -0.4071  -0.2738   2.8363
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -7.927497   0.846875  -9.361  < 2e-16 ***
## gender1          0.422202   0.133313   3.167 0.001540 **
## age              0.066797   0.008110   8.237  < 2e-16 ***
## education2      -0.079672   0.146967  -0.542 0.587743
## education3      -0.329631   0.183167  -1.800 0.071921 .
## education4      -0.236143   0.213615  -1.105 0.268960
## cigsPerDay       0.020000   0.005146   3.886 0.000102 ***
## BPMeds1         -0.002423   0.294477  -0.008 0.993434
## prevalentStroke1  1.152421  0.659094   1.748 0.080379 .
## prevalentHyp1    0.338398   0.166699   2.030 0.042358 *
## diabetes1       -0.005002   0.374594  -0.013 0.989345
## totChol          0.003606   0.001338   2.696 0.007017 **
## sysBP            0.014442   0.004495   3.213 0.001315 **
## diaBP           -0.007077   0.007813  -0.906 0.365014
## BMI              0.011682   0.015070   0.775 0.438211
## heartRate       -0.011470   0.005157  -2.224 0.026137 *
## glucose          0.007397   0.002634   2.808 0.004983 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2168.1  on 2560  degrees of freedom
## Residual deviance: 1894.3  on 2544  degrees of freedom
## AIC: 1928.3
##
## Number of Fisher Scoring iterations: 5
```

```
exp(coef(fitGLMQ5))
```

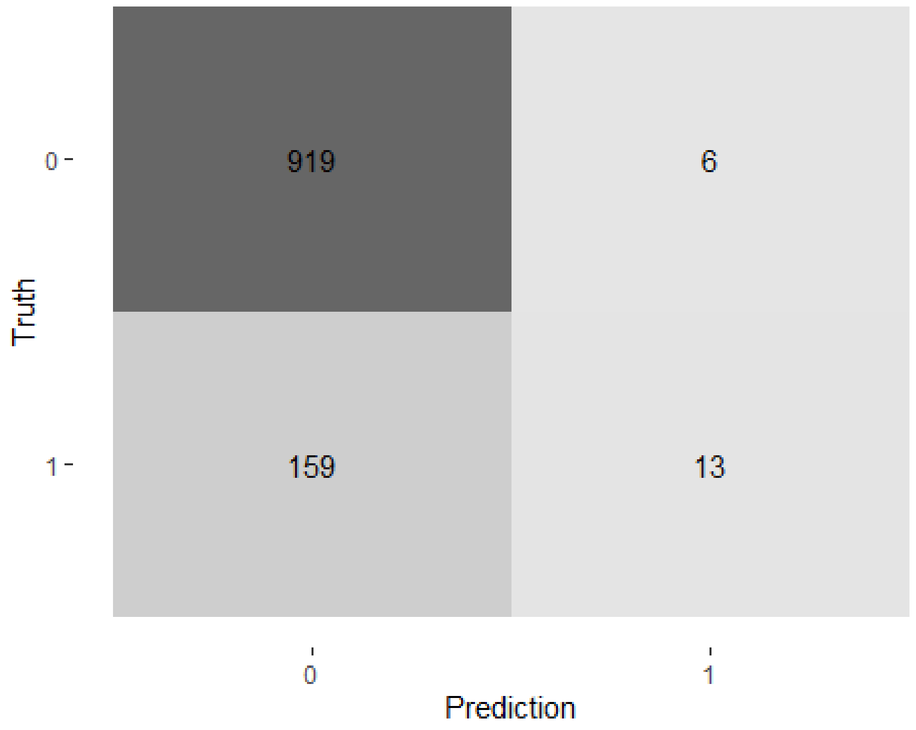
```
##      (Intercept)          gender1              age       education2
##     0.0003606879     1.5253171095     1.0690784440     0.9234189417
##       education3       education4       cigsPerDay          BPMeds1
##     0.7191887265     0.7896676736     1.0202012574     0.9975796686
## prevalentStroke1    prevalentHyp1        diabetes1          totChol
##     3.1658488040     1.4026980839     0.9950101842     1.0036127972
##            sysBP            diaBP              BMI        heartRate
##     1.0145465769     0.9929479273     1.0117507851     0.9885958031
##          glucose
##     1.0074239785
```

Question 6

```
resultsLog  <-
    glm(TenYearCHD ~. -currentSmoker, family = binomial(), data= dffTrain )
%>%
    predict(dffTest, type= 'response') %>%
    bind_cols(dffTest, predictedProb=.) %>%
    mutate(predictedClass = as.factor(ifelse(predictedProb > 0.5, 1, 0)))
resultsLog

## # A tibble: 1,097 x 18
##    gender   age education currentSmoker cigsPerDay BPMeds prevalentStroke
##    <fct>  <dbl> <fct>     <fct>                 <dbl> <fct>  <fct>
##  1 1         48 1         1                        20 0      0
##  2 0         43 2         0                         0 0      0
##  3 0         43 2         0                         0 0      0
##  4 0         41 3         0                         0 1      0
##  5 0         52 3         1                        20 0      0
##  6 0         61 3         0                         0 0      0
##  7 1         46 1         1                        20 0      0
##  8 0         63 2         1                        40 0      0
##  9 0         62 1         0                         0 0      0
## 10 1         49 1         1                         2 0      0
## # ... with 1,087 more rows, and 11 more variables: prevalentHyp <fct>,
## #   diabetes <fct>, totChol <dbl>, sysBP <dbl>, diaBP <dbl>, BMI <dbl>,
## #   heartRate <dbl>, glucose <dbl>, TenYearCHD <fct>, predictedProb <dbl>,
## #   predictedClass <fct>

resultsLog %>% group_by(predictedClass ) %>%
  tally() %>%
  mutate(pct = 100*n/sum(n))

## # A tibble: 2 x 3
##   predictedClass     n   pct
##   <fct>          <int> <dbl>
## 1 0               1078 98.3
## 2 1                 19  1.73
```

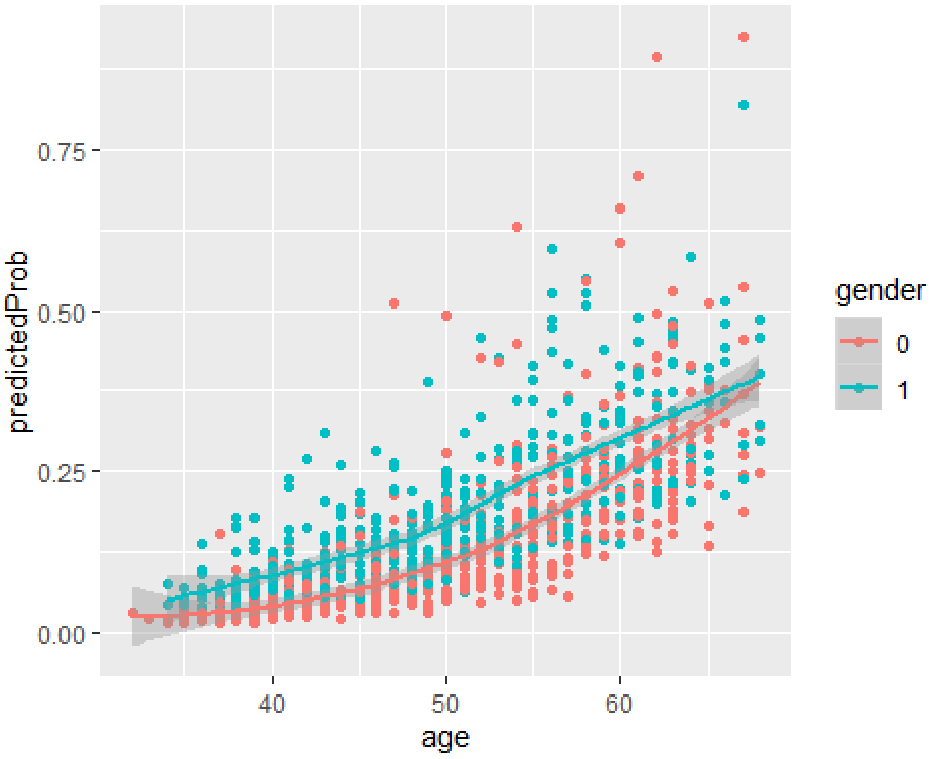
Question 7

```
resultsLog %>%
  conf_mat(truth =TenYearCHD , estimate = predictedClass) %>%
  autoplot(type = 'heatmap')
```

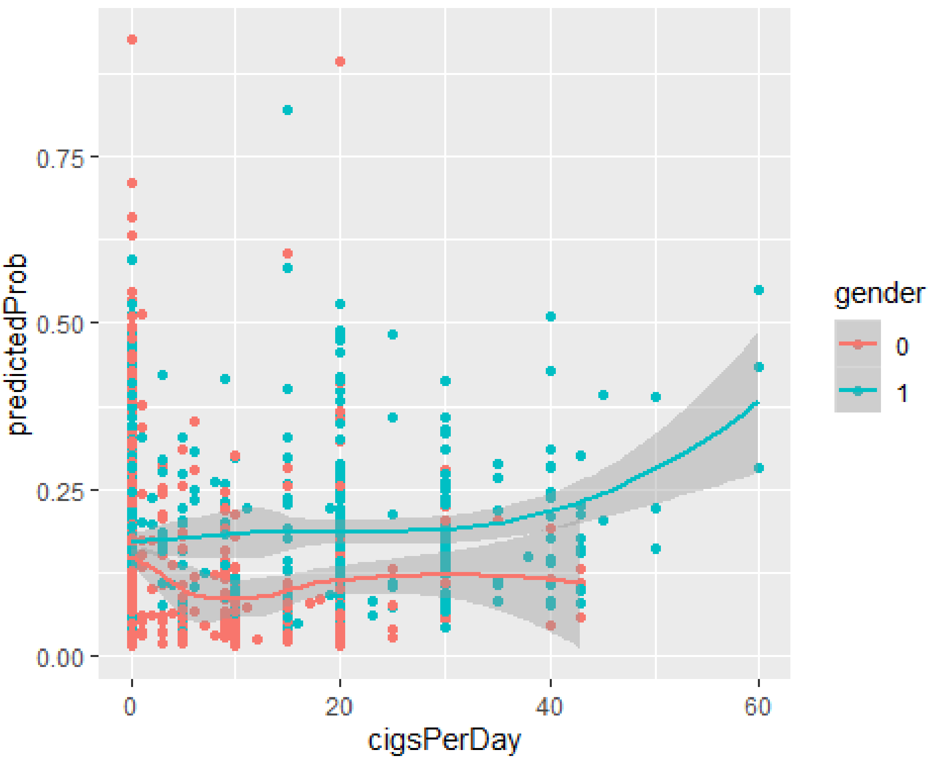


Question 8

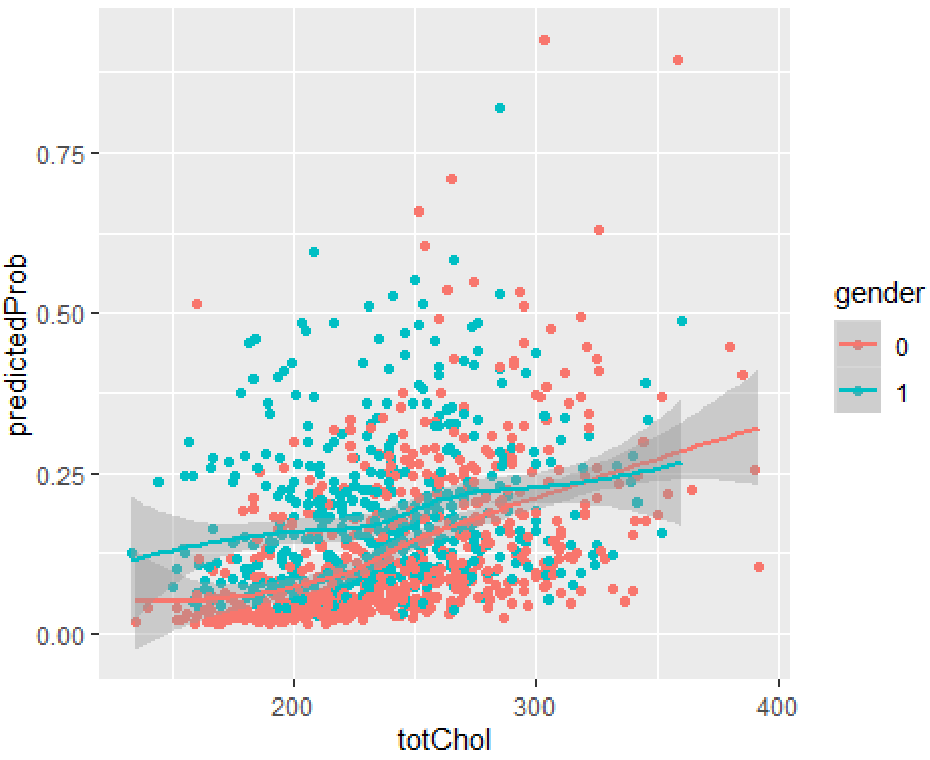
```
plotQ81 <- resultsLog %>%
  ggplot(aes(x= age, y=predictedProb, color=gender)) +
  geom_point() +
  geom_smooth()
plotQ81

## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



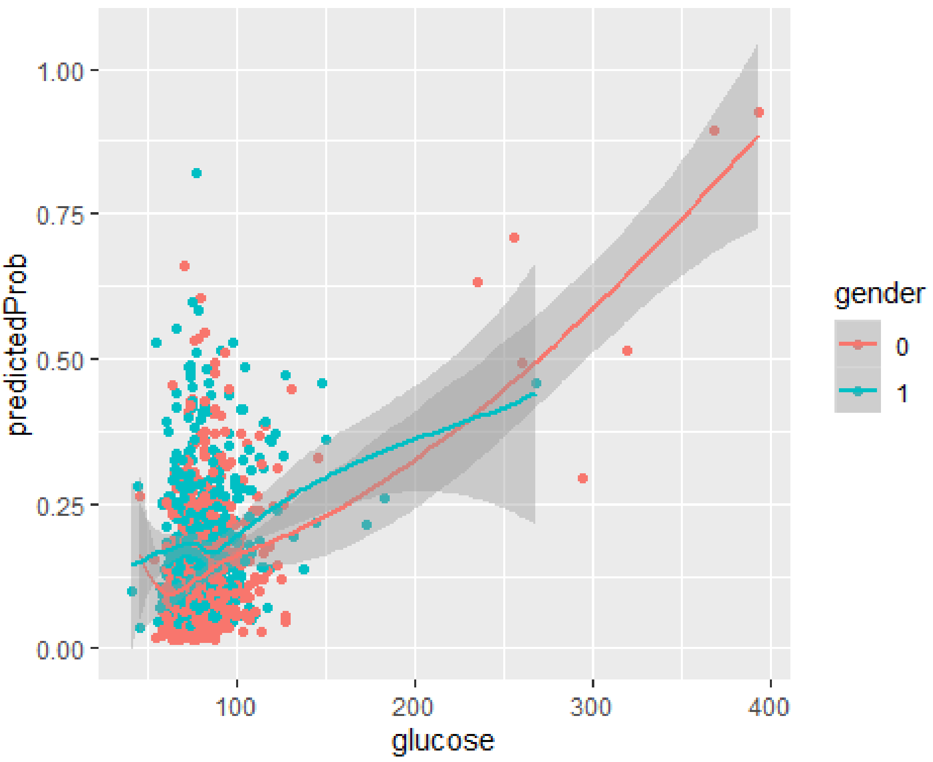
```
plotQ82 <- resultsLog %>%
  ggplot(aes(x= cigsPerDay, y=predictedProb, color=gender)) +
  geom_point()+
  geom_smooth()
plotQ82

## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



```
plotQ83 <- resultsLog %>%
  ggplot(aes(x= totChol, y=predictedProb, color=gender)) +
  geom_point() +
  geom_smooth()
plotQ83

## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



```
plotQ84 <- resultsLog %>%
  ggplot(aes(x= glucose, y=predictedProb,color=gender)) +
  geom_point() +
  geom_smooth()
plotQ84

## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Question 9

```
library(e1071)
resultsLogCaret  <-
    train(TenYearCHD ~. -currentSmoker, family = 'binomial', data= dffTrain,
method= 'glm' ) %>%
    predict(dffTest, type= 'raw') %>%
    bind_cols(dffTest, predictedClass=.)
resultsLogCaret %>%
  xtabs(~predictedClass+TenYearCHD, .) %>%
  confusionMatrix(positive = '1')

## Confusion Matrix and Statistics
##
##                 TenYearCHD
## predictedClass   0   1
##               0 919 159
##               1   6  13
##
##                Accuracy : 0.8496
##                  95% CI : (0.827, 0.8702)
##     No Information Rate : 0.8432
##     P-Value [Acc > NIR] : 0.297
##
##                   Kappa : 0.1083
##
##  Mcnemar's Test P-Value : <2e-16
```

```
##
##             Sensitivity : 0.07558
##             Specificity : 0.99351
##          Pos Pred Value : 0.68421
##          Neg Pred Value : 0.85250
##              Prevalence : 0.15679
##          Detection Rate : 0.01185
##    Detection Prevalence : 0.01732
##       Balanced Accuracy : 0.53455
##
##        'Positive' Class : 1
##
```