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.

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

library("tidyverse")

## ── Attaching packages ─────────────────────────────────────── tidyverse
1.3.0 ──

## ✓ ggplot2 3.2.1     ✓ purrr   0.3.3
## ✓ tibble  2.1.3     ✓ dplyr   0.8.3
## ✓ tidyr   1.0.2     ✓ stringr 1.4.0
## ✓ readr   1.3.1     ✓ forcats 0.4.0

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

library("tidymodels")

## ── Attaching packages ─────────────────────────────────────── tidymodels
0.0.3 ──

## ✓ broom     0.5.4     ✓ recipes   0.1.9
## ✓ dials     0.0.4     ✓ rsample   0.0.5
## ✓ infer     0.5.1     ✓ yardstick 0.0.5
## ✓ 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()
```

```
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
```

```
library("skimr")
```

```
#install.packages("caret")
```

```
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
```

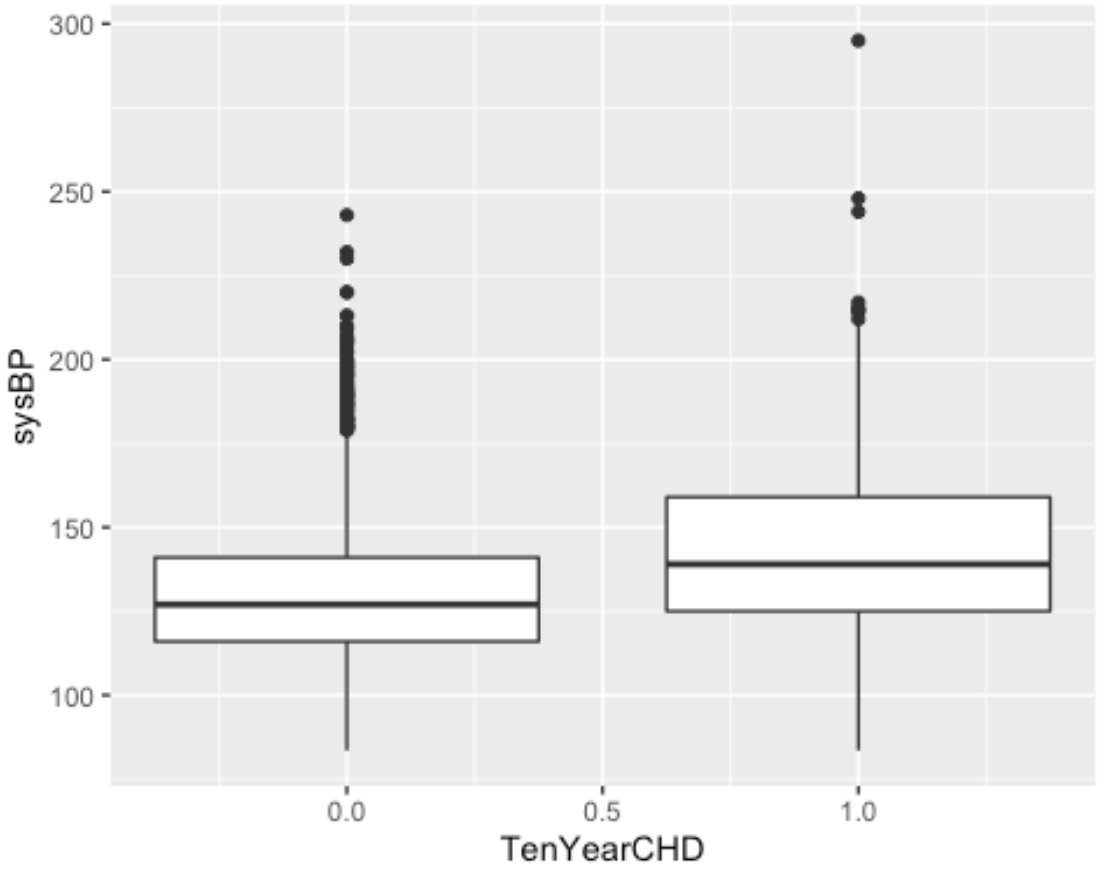
```
dff <-

read_csv("/Users/shruthinair/Desktop/Lumos/DM/Data/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(),
```

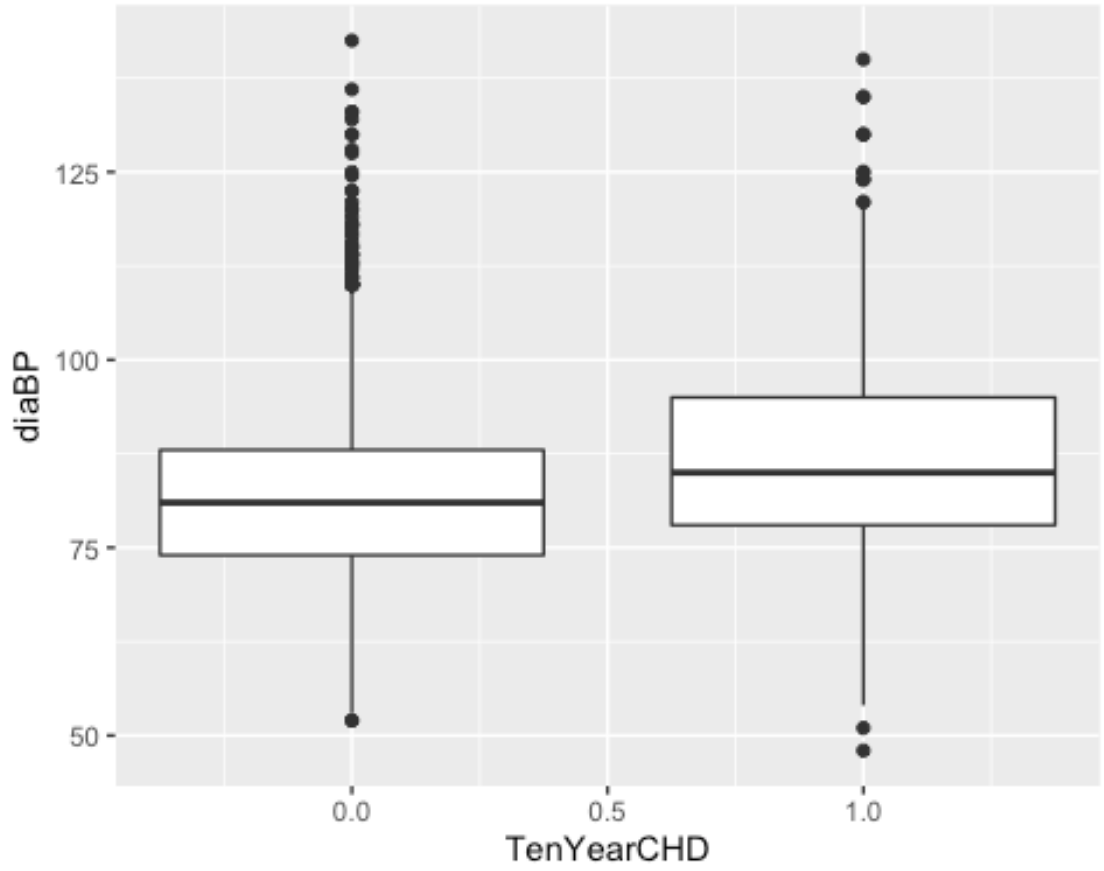
```
##   diaBP = col_double(),
##   BMI = col_double(),
##   heartRate = col_double(),
##   glucose = col_double(),
##   TenYearCHD = col_double()
## )

colsToFactor <- c('gender', 'education', 'currentSmoker', 'BPMeds',
'prevalentStroke', 'prevalentHyp', 'diabetes')
dff <- dff %>%
  mutate_at(colsToFactor, ~factor(.))
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:

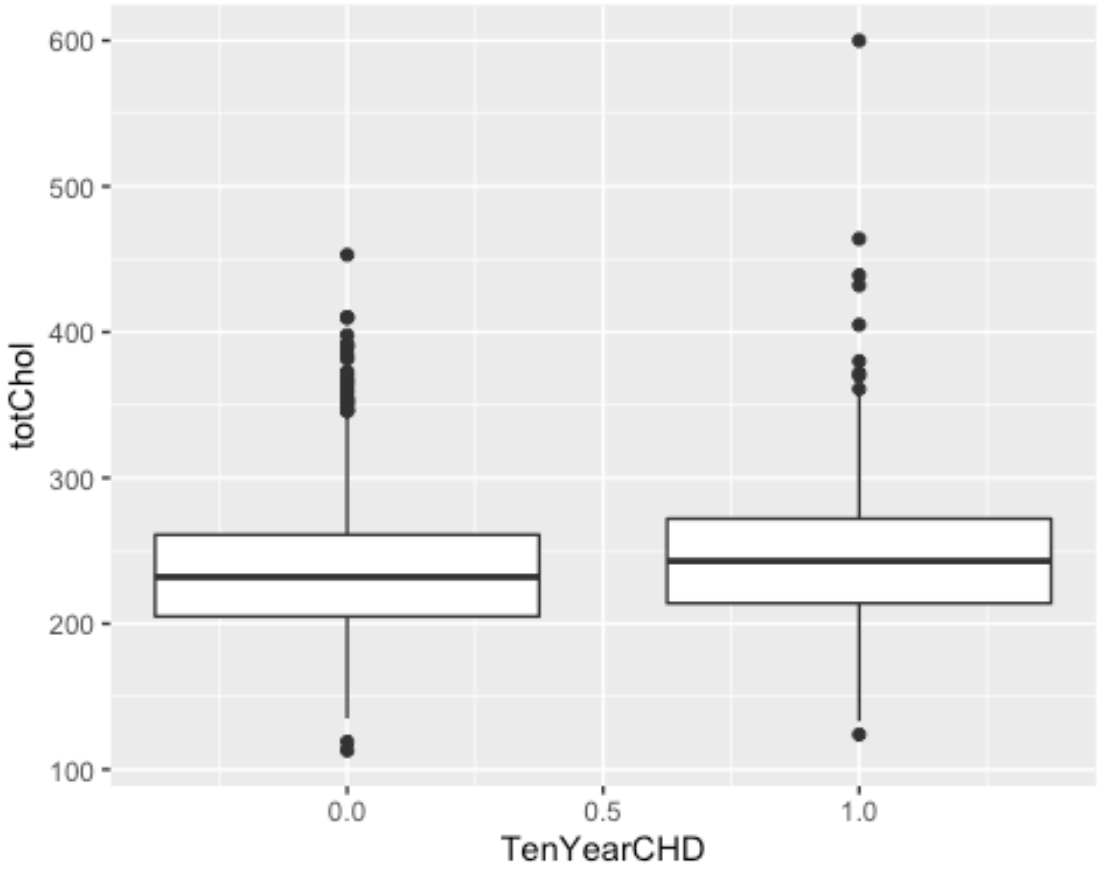
```
boxplot <-
  dff %>%
  ggplot(aes(x = TenYearCHD, y = sysBP)) + geom_boxplot(aes(group =
TenYearCHD))
boxplot
```




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




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



Question 2i -

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

Question 2ii -

```
dffTrain %>%
  group_by(gender) %>%
  tally() %>%
  mutate(prop = n/sum(n))
```

```
## # A tibble: 2 x 3
##   gender     n  prop
##   <fct>  <int> <dbl>
## 1 0       1419 0.554
## 2 1       1142 0.446
```

```
dffTest %>%
  group_by(gender) %>%
  tally() %>%
  mutate(prop = n/sum(n))
```

```
## # A tibble: 2 x 3
##   gender     n  prop
##   <fct>  <int> <dbl>
## 1 0        616 0.562
## 2 1        481 0.438
```

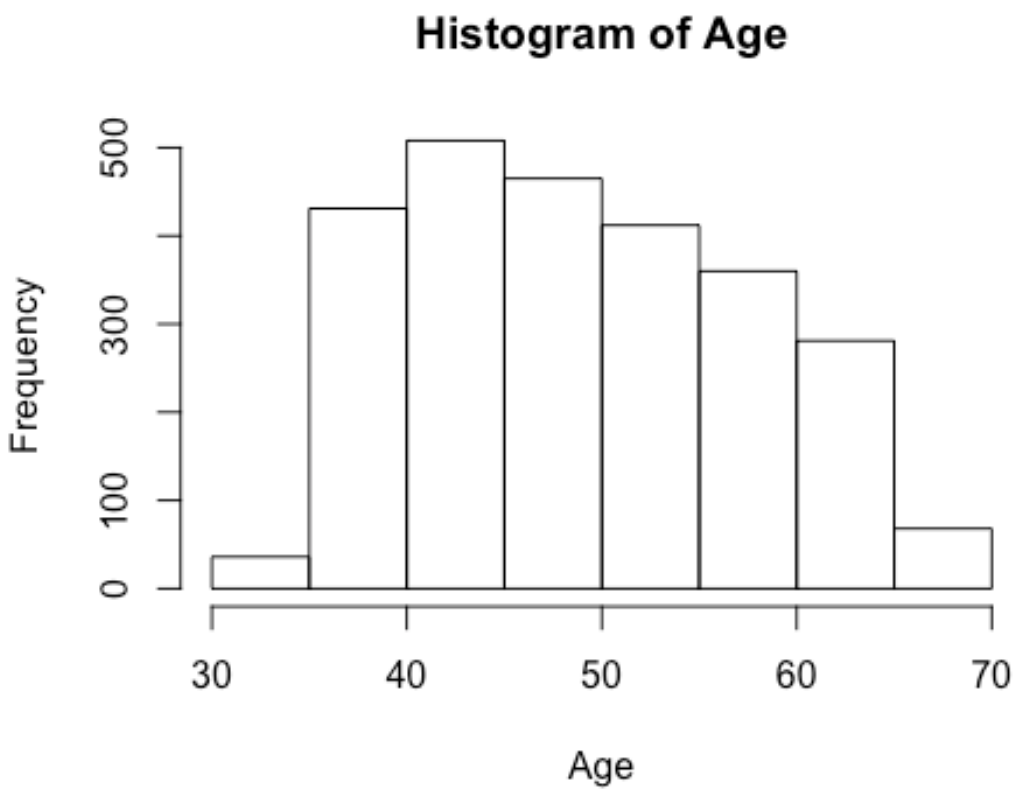
```
dffTrain %>%
  group_by(ageGroup = cut_interval (age, length =10)) %>%
  tally() %>%
  mutate(prop = n/sum(n))
```

```
## # A tibble: 4 x 3
##   ageGroup     n  prop
##   <fct>    <int> <dbl>
## 1 [30,40]    467 0.182
## 2 (40,50]    973 0.380
## 3 (50,60]    772 0.301
## 4 (60,70]    349 0.136
```

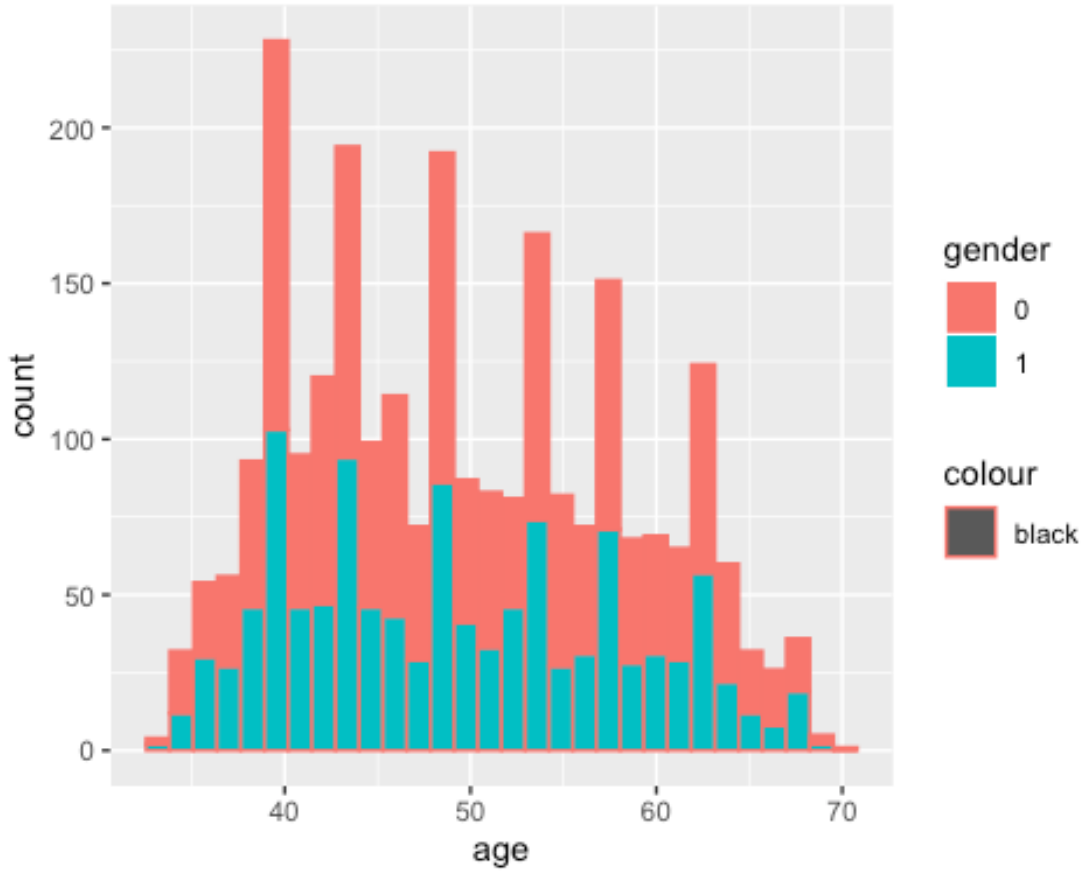
```
dffTest %>%
  group_by(ageGroup = cut_interval (age, length =10)) %>%
  tally() %>%
  mutate(pct = n/sum(n))
```

```
## # A tibble: 4 x 3
##   ageGroup     n   pct
##   <fct>    <int> <dbl>
## 1 [30,40]    181 0.165
## 2 (40,50]    421 0.384
## 3 (50,60]    346 0.315
## 4 (60,70]    149 0.136
```

```
Age <- dffTrain$age
H <-hist(Age)
```



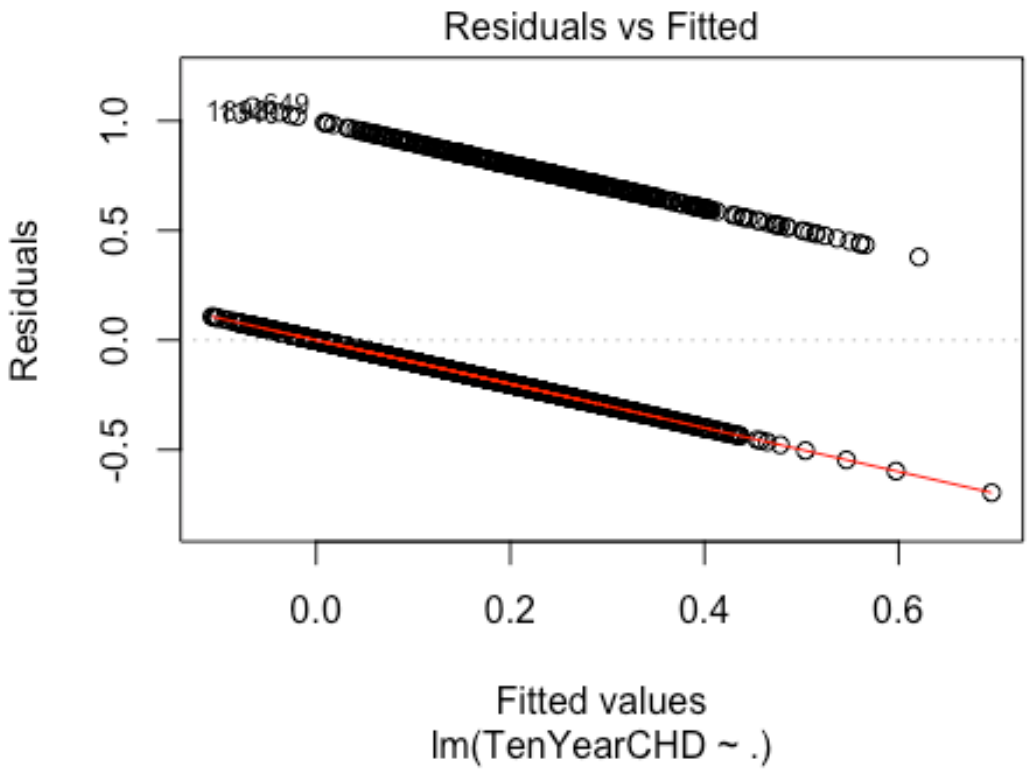

```
ggplot(data = dffTrain, aes(x= age, fill = gender, color = 'black')) +
geom_histogram()

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

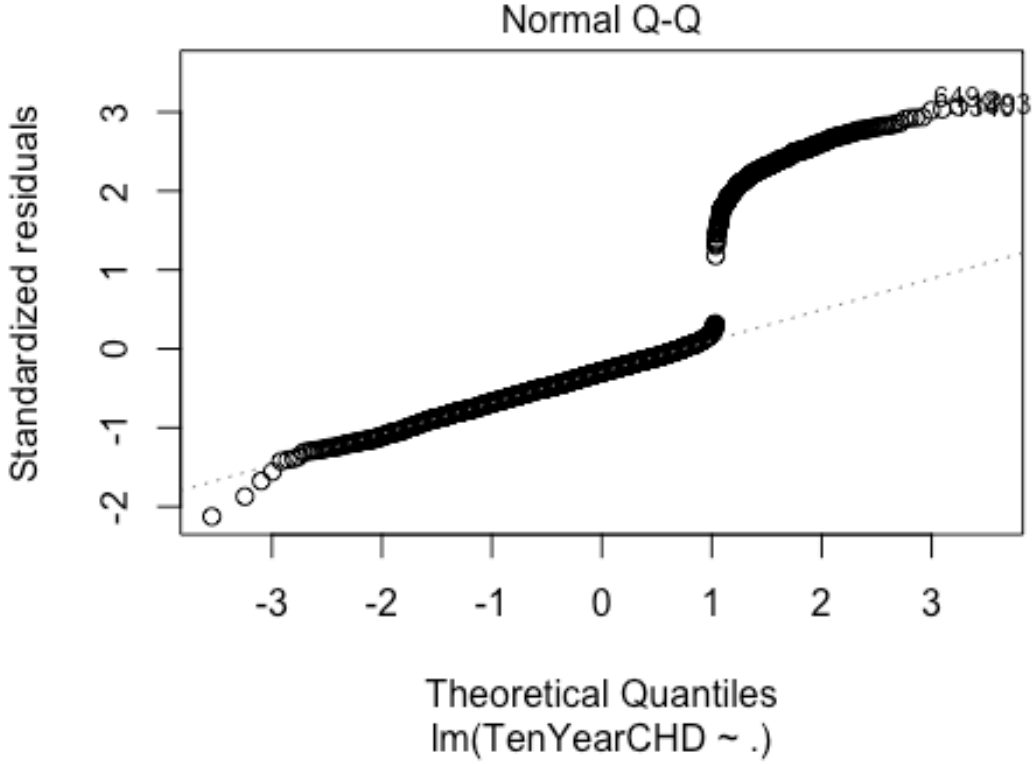


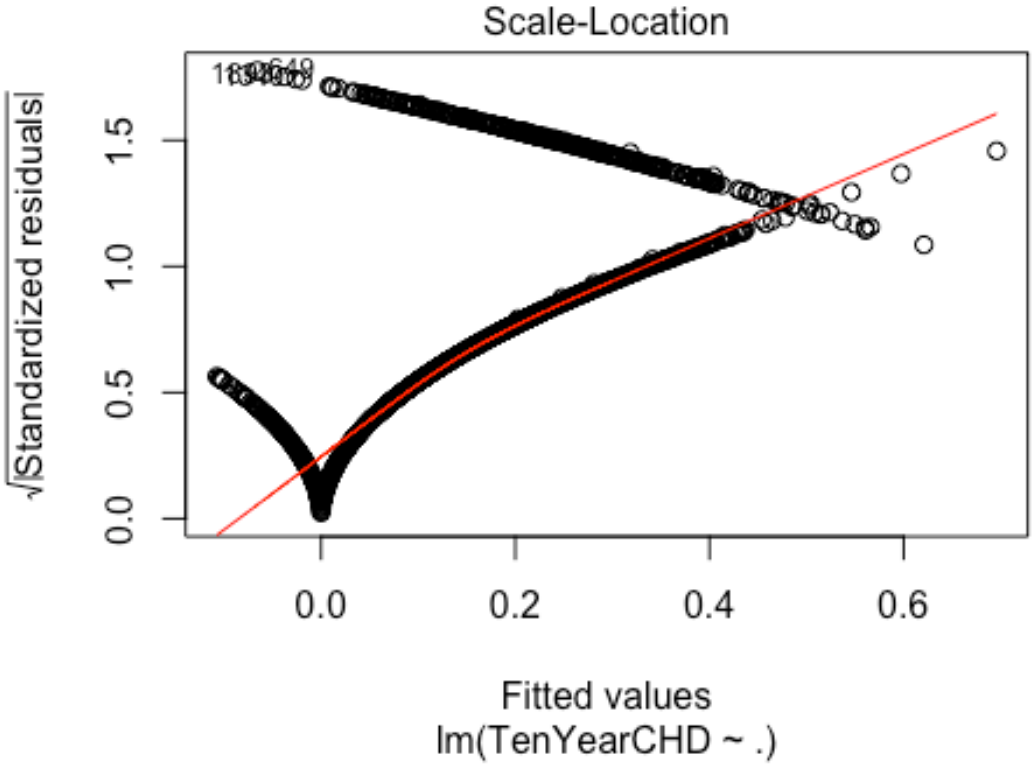
Question 3:

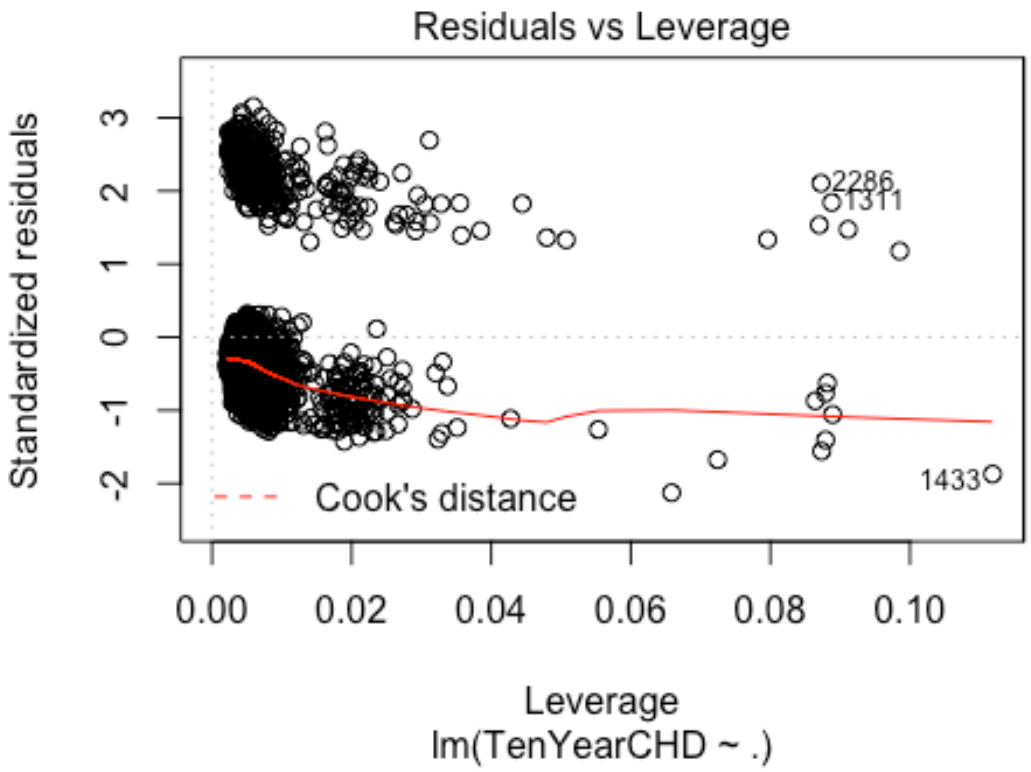
```
fitLPM <-
  lm(formula = TenYearCHD ~ ., data = dffTrain)
summary(fitLPM)

##
## 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

plot(fitLPM)
```


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


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


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
0.0
0.2
0.4
0.6
Fitted values
lm(TenYearCHD ~ .)



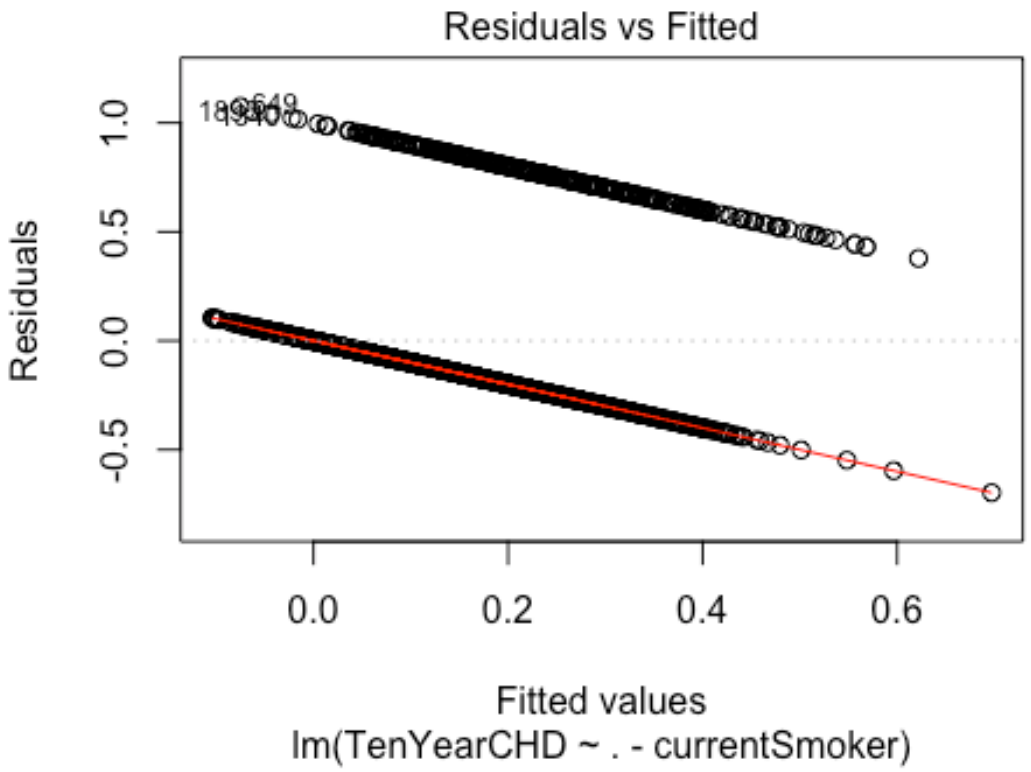

```
fitLPM <-
  lm(formula = TenYearCHD ~ .-currentSmoker, data = dffTrain)
summary(fitLPM)

##
## 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 *
```

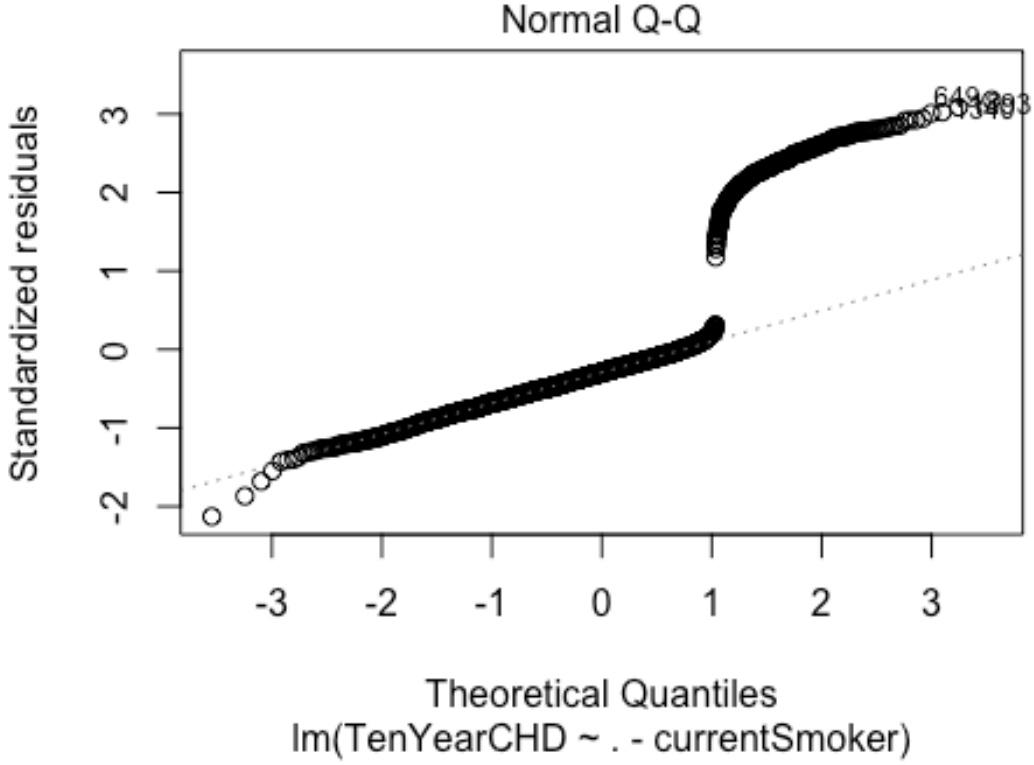
```
## prevalentHyp1     0.0448426  0.0208855   2.147 0.031882 *
## diabetes1         0.0203925  0.0513670   0.397 0.691403
## totChol           0.0002875  0.0001590   1.809 0.070633 .
## sysBP             0.0023882  0.0005798   4.119 3.92e-05 ***
## diaBP            -0.0016833  0.0009708  -1.734 0.083051 .
## 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
```

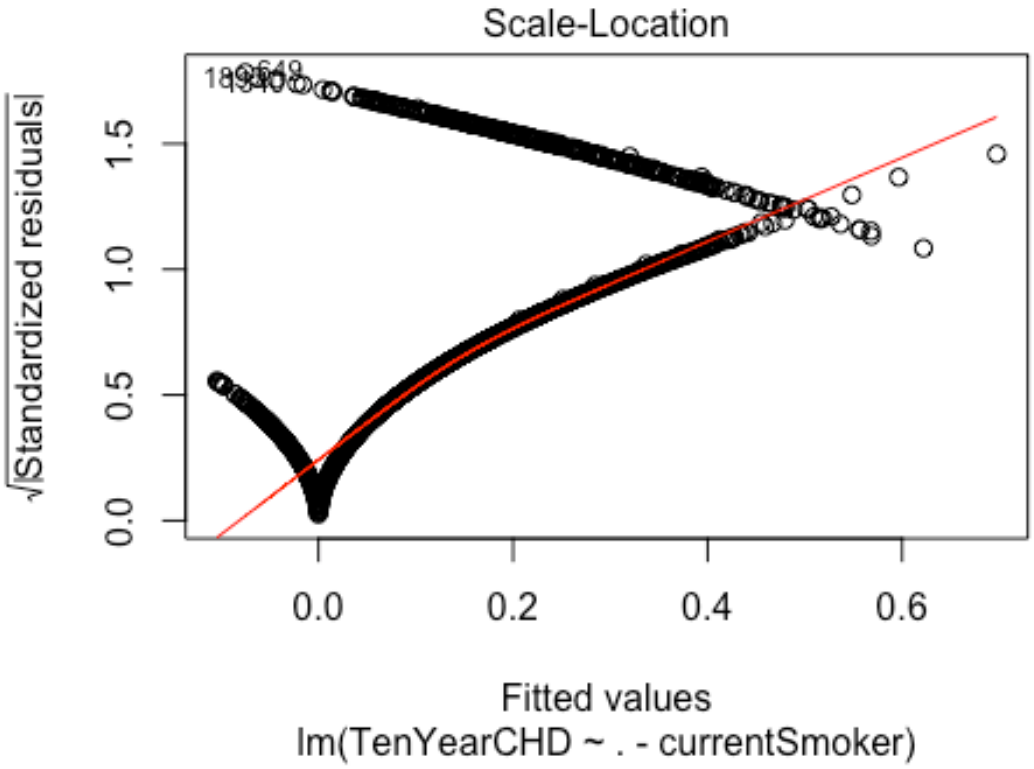
```
car::vif(fitLPM)
```

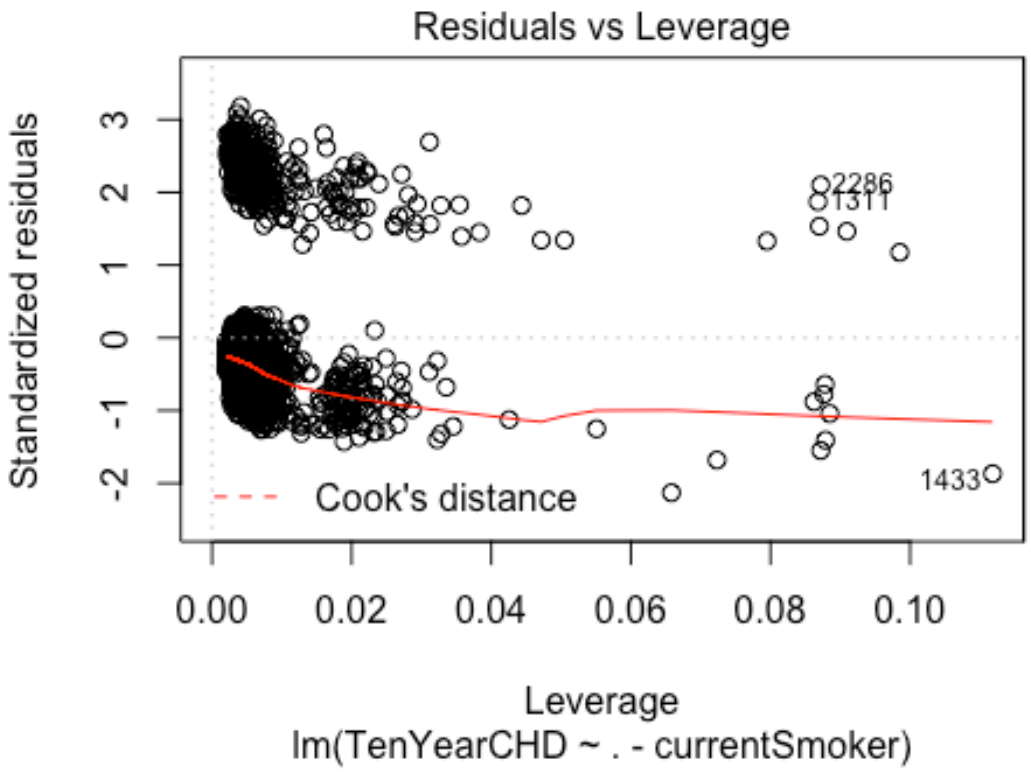
```
##                     GVIF Df GVIF^(1/(2*Df))
## gender          1.227561  1        1.107954
## age             1.390293  1        1.179107
## education       1.139163  3        1.021953
## cigsPerDay      1.202282  1        1.096486
## BPMeds          1.106788  1        1.052040
## prevalentStroke 1.006566  1        1.003278
## prevalentHyp    2.057379  1        1.434357
## diabetes        1.630611  1        1.276954
## totChol         1.106882  1        1.052085
## sysBP           3.777149  1        1.943489
## diaBP           2.993948  1        1.730303
## BMI             1.218397  1        1.103810
## heartRate       1.095825  1        1.046817
## glucose         1.645572  1        1.282799
```

```
plot(fitLPM)
```


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


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


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
0.0
0.2
0.4
0.6
Fitted values
lm(TenYearCHD ~ . - currentSmoker)



Question 4:

```
resultsLPM <-
    #lm(formula = TenYearCHD ~ .-currentSmoker, data = dffTrain) %>%
  fitLPM %>%
    predict(dffTest, type='response') %>%
  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()
```

```
## # A tibble: 2 x 2
##   TenYearCHD     n
##        <dbl> <int>
## 1          0   925
## 2          1   172
```

```
resultsLPM %>%
  group_by(predictedClass) %>%
  tally()
```

```
## # A tibble: 2 x 2
##   predictedClass     n
##            <dbl> <int>
## 1              0  1087
## 2              1    10
```

```
dffTrain$TenYearCHD <- as.factor(dffTrain$TenYearCHD)
dffTest$TenYearCHD <- as.factor(dffTest$TenYearCHD)
```

Question 5:

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

```
##
## 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(coefficients(fitLog))
```

```
##      (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:

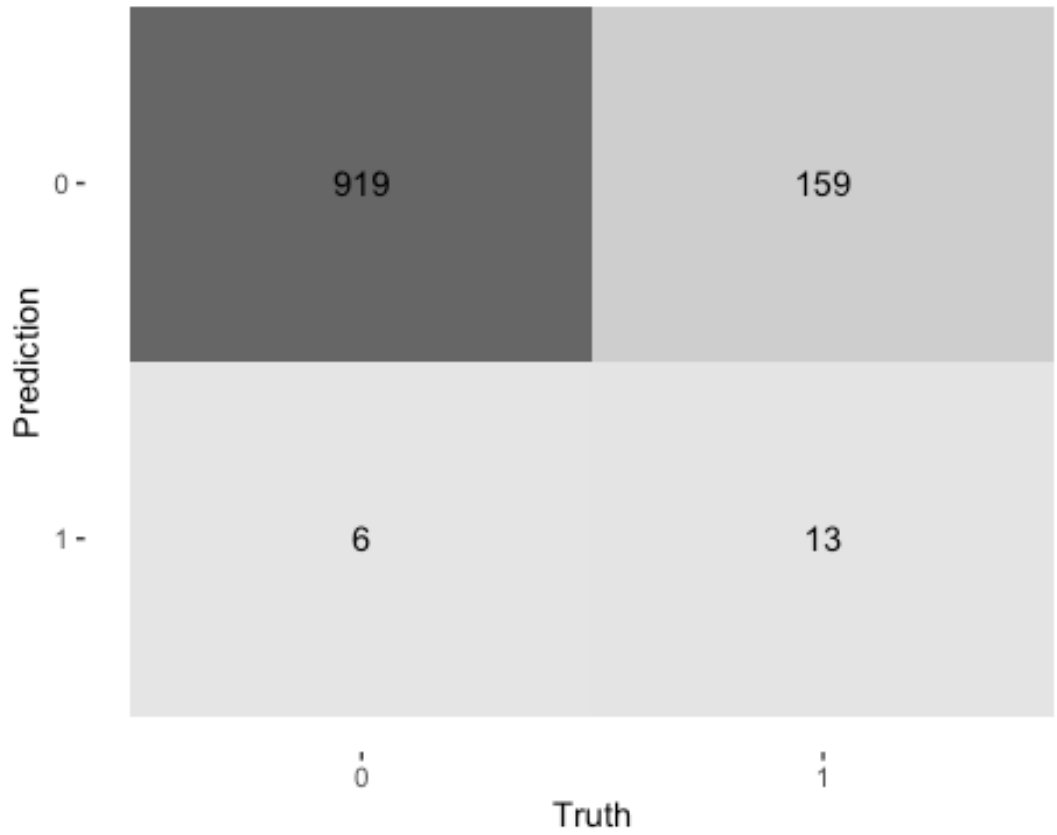
```
resultsLog <-
    #glm(formula = TenYearCHD ~ .-currentSmoker, data = dffTrain) %>%
  fitLog %>%
    predict(dffTest, type='response') %>%
  bind_cols(dffTest, predictedProb=.) %>%
    mutate(predictedClass = as.factor(ifelse (predictedProb>0.5,1,0)))

resultsLog %>%
  group_by(predictedClass) %>%
  tally()
```

```
## # A tibble: 2 x 2
##   predictedClass     n
##   <fct>          <int>
```

```
## 1 0                1078
## 2 1                  19
```

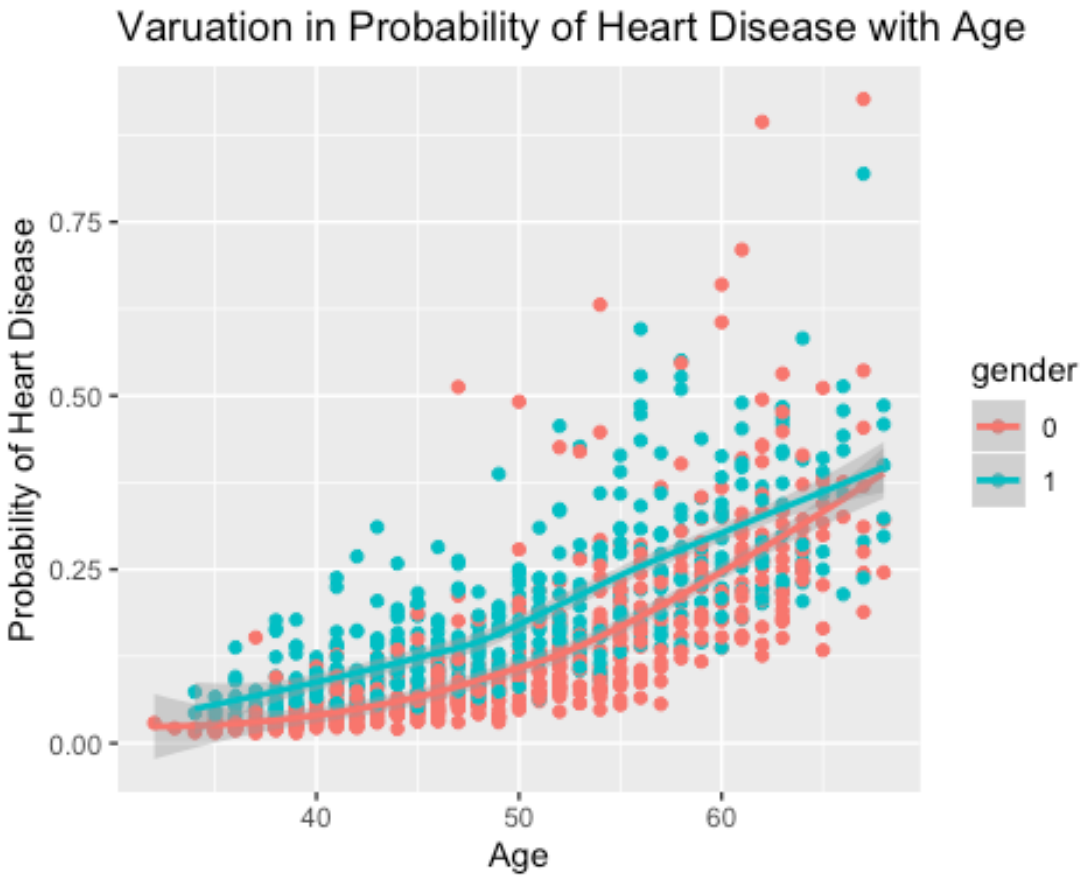
Question 7:

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

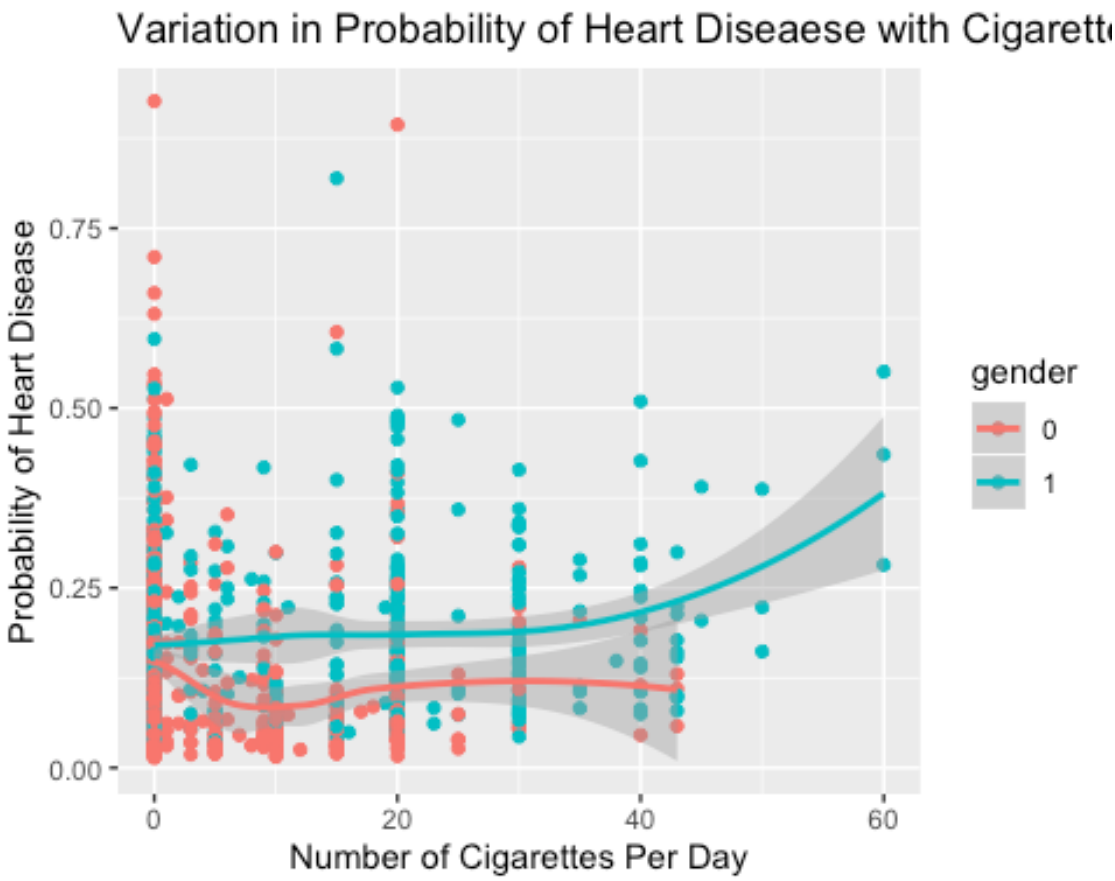


Question 8:

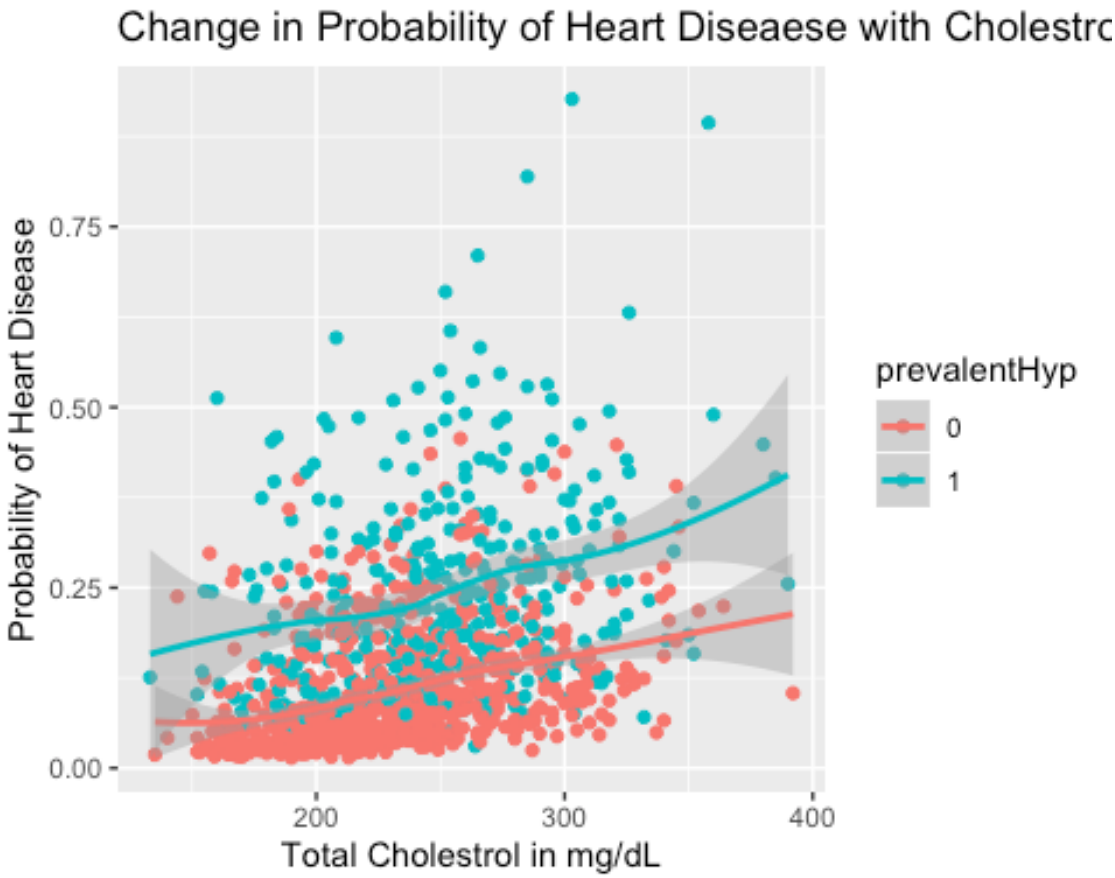
```
ggplot(data=resultsLog, aes(x=age, y=predictedProb, color=gender)) + labs(x=
"Age", y= "Probability of Heart Disease") +ggtitle("Varuation in Probability
of Heart Disease with Age") + geom_point() + geom_smooth()

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




```
ggplot(data=resultsLog, aes(x=cigsPerDay, y=predictedProb, color = gender)) +
labs(x= "Number of Cigarettes Per Day", y= "Probability of Heart Disease") +
ggtitle("Variation in Probability of Heart Diseaese with Cigarettes Smoked
Per Day") + geom_point() + geom_smooth()

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




```
ggplot(data=resultsLog, aes(x=totChol, y=predictedProb, color =
prevalentHyp)) + labs(x= "Total Cholestrol in mg/dL", y= "Probability of
Heart Disease") + ggtitle("Change in Probability of Heart Diseaese with
Cholestrol Levels") + geom_point() + geom_smooth()

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




```
ggplot(data=resultsLog, aes(x=glucose, y=predictedProb, color = gender)) +
labs(x= "Glucose Level in mg/dL", y= "Probability of Heart Disease")
+ggtitle("Change in Probability of Heart Diseaese with Age") + geom_point() +
geom_smooth()

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


```
#install.packages('e1071', dependencies=TRUE)
```

Question 9:

```
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
##
```

Question 10:

```
dfb <-
  read_csv("/Users/shruthinair/Desktop/Lumos/DM/Data/lab3BancoPortugal.csv")

## Parsed with column specification:
## cols(
##   .default = col_double(),
##   job = col_character(),
##   marital = col_character(),
##   education = col_character(),
##   default = col_character(),
##   housing = col_character(),
##   loan = col_character(),
##   contact = col_character(),
##   month = col_character(),
##   day_of_week = col_character(),
##   poutcome = col_character(),
##   agegroup = col_character()
## )

## See spec(...) for full column specifications.

colsToFactorB <- c('newcustomer', 'agegroup', 'job', 'marital', 'education',
'default', 'housing', 'loan', 'contact', 'month', 'day_of_week', 'poutcome')
dfb <- dfb %>%
  mutate_at(colsToFactorB, ~factor(.))
str(dfb)

## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 30488 obs. of  23
variables:
##  $ age           : num  56 37 40 56 59 24 25 25 29 57 ...
##  $ job           : Factor w/ 11 levels "admin.","blue-collar",..: 4 8 1 8
```

```
1 10 8 8 2 4 ...
##  $ marital       : Factor w/ 3 levels "divorced","married",..: 2 2 2 2 2 3
3 3 3 1 ...
##  $ education     : Factor w/ 7 levels "basic.4y","basic.6y",..: 1 4 2 4 6
6 4 4 4 1 ...
##  $ default       : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 ...
##  $ housing       : Factor w/ 2 levels "no","yes": 1 2 1 1 1 2 2 2 1 2 ...
##  $ loan          : Factor w/ 2 levels "no","yes": 1 1 1 2 1 1 1 1 2 1 ...
##  $ contact       : Factor w/ 2 levels "cellular","telephone": 2 2 2 2 2 2
2 2 2 2 ...
##  $ month         : Factor w/ 10 levels "apr","aug","dec",..: 7 7 7 7 7 7 7
7 7 7 ...
##  $ day_of_week   : Factor w/ 5 levels "fri","mon","thu",..: 2 2 2 2 2 2 2
2 2 2 ...
##  $ duration      : num  261 226 151 307 139 380 50 222 137 293 ...
##  $ campaign      : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ pdays         : num  999 999 999 999 999 999 999 999 999 999 ...
##  $ previous      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ poutcome      : Factor w/ 3 levels "failure","nonexistent",..: 2 2 2 2
2 2 2 2 2 2 ...
##  $ emp.var.rate  : num  1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 ...
##  $ cons.price.idx: num  94 94 94 94 94 ...
##  $ cons.conf.idx : num  -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -
36.4 -36.4 ...
##  $ euribor3m     : num  4.86 4.86 4.86 4.86 4.86 ...
##  $ nr.employed   : num  5191 5191 5191 5191 5191 ...
##  $ openedAccount : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ agegroup      : Factor w/ 4 levels "Adults","Senior Citizens",..: 1 1 1
1 1 4 4 4 4 1 ...
##  $ newcustomer   : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...

set.seed(123)
dfbTrain <- dfb %>% sample_frac(0.7)
dfbTest <- setdiff(dfb, dfbTrain)

fitbLM <-
  lm(formula = openedAccount ~ .-newcustomer, data = dfbTrain)
summary(fitbLM)

##
## Call:
## lm(formula = openedAccount ~ . - newcustomer, data = dfbTrain)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.10213 -0.10923 -0.01913  0.03401  1.14213
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -2.362e+01  4.198e+00  -5.626 1.86e-08 ***
```

```
## age                             3.143e-04  3.468e-04   0.906 0.364779
## jobblue-collar                 -9.368e-03  7.018e-03  -1.335 0.181939
## jobentrepreneur                -4.490e-03  1.044e-02  -0.430 0.667148
## jobhousemaid                   -3.886e-03  1.362e-02  -0.285 0.775421
## jobmanagement                   4.381e-03  7.731e-03   0.567 0.570918
## jobretired                      1.164e-02  1.288e-02   0.903 0.366308
## jobself-employed               -9.021e-03  1.049e-02  -0.860 0.390022
## jobservices                    -1.109e-02  7.366e-03  -1.505 0.132274
## jobstudent                      4.405e-02  1.436e-02   3.067 0.002164 **
## jobtechnician                   6.380e-03  6.296e-03   1.013 0.310876
## jobunemployed                   6.012e-03  1.229e-02   0.489 0.624659
## maritalmarried                 -2.431e-03  5.985e-03  -0.406 0.684580
## maritalsingle                   1.128e-03  6.842e-03   0.165 0.869001
## educationbasic.6y               3.831e-03  1.097e-02   0.349 0.726985
## educationbasic.9y              -6.213e-03  8.528e-03  -0.728 0.466318
## educationhigh.school            3.014e-03  8.681e-03   0.347 0.728418
## educationilliterate             1.525e-01  8.967e-02   1.700 0.089095 .
## educationprofessional.course    7.195e-03  9.518e-03   0.756 0.449707
## educationuniversity.degree      1.502e-02  8.827e-03   1.701 0.088868 .
## defaultyes                     -9.412e-03  2.679e-01  -0.035 0.971973
## housingyes                      1.120e-03  3.709e-03   0.302 0.762785
## loanyes                        -4.857e-03  5.038e-03  -0.964 0.334978
## contacttelephone               -5.716e-02  6.923e-03  -8.257  < 2e-16 ***
## monthaug                        1.005e-01  1.547e-02   6.496 8.45e-11 ***
## monthdec                        8.779e-02  2.919e-02   3.007 0.002639 **
## monthjul                        1.920e-02  9.913e-03   1.937 0.052753 .
## monthjun                       -6.133e-02  1.529e-02  -4.013 6.03e-05 ***
## monthmar                        2.685e-01  1.874e-02  14.333  < 2e-16 ***
## monthmay                       -3.823e-02  9.214e-03  -4.150 3.34e-05 ***
## monthnov                       -2.745e-02  1.186e-02  -2.314 0.020670 *
## monthoct                        3.168e-02  1.816e-02   1.744 0.081108 .
## monthsep                        4.784e-02  2.218e-02   2.157 0.031019 *
## day_of_weekmon                 -1.087e-02  5.892e-03  -1.846 0.064971 .
## day_of_weekthu                  4.042e-03  5.842e-03   0.692 0.489059
## day_of_weektue                  1.446e-02  5.961e-03   2.426 0.015265 *
## day_of_weekwed                  1.631e-02  5.923e-03   2.753 0.005903 **
## duration                        4.721e-04  7.113e-06  66.376  < 2e-16 ***
## campaign                        8.685e-04  6.974e-04   1.245 0.212987
## pdays                          -1.557e-04  3.303e-05  -4.714 2.44e-06 ***
## previous                       -8.929e-03  8.528e-03  -1.047 0.295101
## poutcomenonexistent             4.297e-02  1.146e-02   3.751 0.000177 ***
## poutcomesuccess                 1.625e-01  3.259e-02   4.985 6.23e-07 ***
## emp.var.rate                   -1.868e-01  1.667e-02 -11.206  < 2e-16 ***
## cons.price.idx                  2.433e-01  2.798e-02   8.695  < 2e-16 ***
## cons.conf.idx                   3.502e-03  9.614e-04   3.643 0.000270 ***
## euribor3m                       5.696e-02  1.402e-02   4.063 4.86e-05 ***
## nr.employed                     1.739e-04  3.354e-04   0.519 0.604112
## agegroupSenior Citizens         3.381e-02  1.528e-02   2.212 0.026954 *
## agegroupTeenagers               1.722e-01  5.707e-02   3.017 0.002552 **
## agegroupYoung Adults            1.502e-02  6.006e-03   2.502 0.012374 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2678 on 21291 degrees of freedom
## Multiple R-squared:  0.3496, Adjusted R-squared:  0.3481
## F-statistic: 228.9 on 50 and 21291 DF,  p-value: < 2.2e-16
```

```
car::vif(fitbLM)
```

```
##                       GVIF Df GVIF^(1/(2*Df))
## age               3.877015  1        1.969014
## job               6.275617 10        1.096182
## marital           1.337018  2        1.075312
## education         3.242908  6        1.103006
## default           1.000900  1        1.000450
## housing           1.015811  1        1.007874
## loan              1.004258  1        1.002127
## contact           3.146174  1        1.773746
## month           184.285688  9        1.336167
## day_of_week       1.048170  4        1.005898
## duration          1.017215  1        1.008571
## campaign          1.052237  1        1.025786
## pdays            12.871371  1        3.587669
## previous          5.802293  1        2.408795
## poutcome         43.248556  2        2.564442
## emp.var.rate    214.155831  1       14.634064
## cons.price.idx   79.364509  1        8.908676
## cons.conf.idx     6.299758  1        2.509932
## euribor3m       184.661677  1       13.589028
## nr.employed     188.044095  1       13.712917
## agegroup          5.125330  3        1.313067
```

```
dfbTrain$openedAccount <- as.factor(dfbTrain$openedAccount)
dfbTest$openedAccount <- as.factor(dfbTest$openedAccount)
```

Model 1:

```
resultsLogCaret1 <-
  train(openedAccount ~ .-duration-newcustomer, family= "binomial", data=
dfbTrain, method= 'glm') %>%
    predict(dfbTest, type='raw') %>%
  bind_cols(dfbTest, predictedClass=.)
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type
== :
## prediction from a rank-deficient fit may be misleading

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type
== :
## prediction from a rank-deficient fit may be misleading
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type
== :
## prediction from a rank-deficient fit may be misleading

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type
== :
## prediction from a rank-deficient fit may be misleading

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type
== :
## prediction from a rank-deficient fit may be misleading
```

```
resultsLogCaret1 %>%
    xtabs (~predictedClass+openedAccount, .) %>%
  confusionMatrix(positive= '1')
```

```
## Confusion Matrix and Statistics
##
##               openedAccount
## predictedClass    0    1
##              0 7833  871
##              1  136  302
##
##                Accuracy : 0.8898
##                  95% CI : (0.8833, 0.8962)
##     No Information Rate : 0.8717
##     P-Value [Acc > NIR] : 6.372e-08
##
##                   Kappa : 0.328
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.25746
##             Specificity : 0.98293
##          Pos Pred Value : 0.68950
##          Neg Pred Value : 0.89993
##              Prevalence : 0.12831
##          Detection Rate : 0.03303
##    Detection Prevalence : 0.04791
##       Balanced Accuracy : 0.62020
##
##        'Positive' Class : 1
##
```

Model 2:

```
resultsLogCaret1 <-
  train(openedAccount ~ agegroup + contact + euribor3m + cons.conf.idx +
cons.price.idx + emp.var.rate + poutcome + pdays + day_of_week + month + job,
family= "binomial", data= dfbTrain, method= 'glm') %>%
    predict(dfbTest, type='raw') %>%
```

```
  bind_cols(dfbTest, predictedClass=.)

resultsLogCaret1 %>%
    xtabs (~predictedClass+openedAccount, .) %>%
  confusionMatrix(positive= '1')

## Confusion Matrix and Statistics
##
##               openedAccount
## predictedClass    0    1
##              0 7838  871
##              1  131  302
##
##                Accuracy : 0.8904
##                  95% CI : (0.8838, 0.8967)
##     No Information Rate : 0.8717
##     P-Value [Acc > NIR] : 2.577e-08
##
##                   Kappa : 0.3297
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.25746
##             Specificity : 0.98356
##          Pos Pred Value : 0.69746
##          Neg Pred Value : 0.89999
##              Prevalence : 0.12831
##          Detection Rate : 0.03303
##    Detection Prevalence : 0.04736
##       Balanced Accuracy : 0.62051
##
##        'Positive' Class : 1
##
```

Model 3:

```
resultsLogCaret2 <-
  train(openedAccount ~ contact + euribor3m + cons.conf.idx + cons.price.idx
+ emp.var.rate + poutcome + pdays + day_of_week + month + job, family=
"binomial", data= dfbTrain, method= 'glm') %>%
    predict(dfbTest, type='raw') %>%
  bind_cols(dfbTest, predictedClass=.)

resultsLogCaret2 %>%
    xtabs (~predictedClass+openedAccount, .) %>%
  confusionMatrix(positive= '1')

## Confusion Matrix and Statistics
##
##               openedAccount
## predictedClass    0    1
```

```
##              0 7839  866
##              1  130  307
##
##                Accuracy : 0.8911
##                  95% CI : (0.8845, 0.8974)
##     No Information Rate : 0.8717
##     P-Value [Acc > NIR] : 8.39e-09
##
##                   Kappa : 0.3351
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.26172
##             Specificity : 0.98369
##          Pos Pred Value : 0.70252
##          Neg Pred Value : 0.90052
##              Prevalence : 0.12831
##          Detection Rate : 0.03358
##    Detection Prevalence : 0.04780
##       Balanced Accuracy : 0.62270
##
##        'Positive' Class : 1
##

resultsLogCaret2

## # A tibble: 9,142 x 24
##      age job   marital education default housing loan  contact month
day_of_week
##    <dbl> <fct> <fct>   <fct>     <fct>   <fct>   <fct> <fct>   <fct> <fct>
##  1    56 hous… married basic.4y  no      no      no    teleph… may   mon
##  2    24 tech… single  professi… no      yes     no    teleph… may   mon
##  3    25 serv… single  high.sch… no      yes     no    teleph… may   mon
##  4    35 blue… married basic.6y  no      yes     no    teleph… may   mon
##  5    32 entr… married high.sch… no      yes     no    teleph… may   mon
##  6    38 admi… single  professi… no      no      no    teleph… may   mon
##  7    35 admi… married universi… no      yes     no    teleph… may   mon
##  8    53 admi… single  professi… no      no      no    teleph… may   mon
##  9    25 tech… single  universi… no      yes     no    teleph… may   mon
## 10    56 admi… married basic.9y  no      yes     no    teleph… may   mon
## # … with 9,132 more rows, and 14 more variables: duration <dbl>,
## #   campaign <dbl>, pdays <dbl>, previous <dbl>, poutcome <fct>,
## #   emp.var.rate <dbl>, cons.price.idx <dbl>, cons.conf.idx <dbl>,
## #   euribor3m <dbl>, nr.employed <dbl>, openedAccount <fct>, agegroup
<fct>,
## #   newcustomer <fct>, predictedClass <fct>
```