Heart Disease Machine Learning Prediction

```
# libraries set up

library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1
```

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

```
library(explore)
```

```
## Warning: package 'explore' was built under R version 4.1.3
```

```
library(ggplot2)
library(ggfortify)
```

```
## Warning: package 'ggfortify' was built under R version 4.1.3
```

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.1.3
```

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

```
## Warning: package 'lattice' was built under R version 4.1.3
```

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

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

```r
library(rpart)
library(e1071)
```

```
## Warning: package ’e1071’ was built under R version 4.1.3
```

```r
library(rattle)
```

```
## Warning: package ’rattle’ was built under R version 4.1.3

## Loading required package: bitops

## Rattle: A free graphical interface for data science with R.
## Version 5.5.1 Copyright (c) 2006-2021 Togaware Pty Ltd.
## Type ’rattle()’ to shake, rattle, and roll your data.
```

```r
library(mlbench)
```

```
## Warning: package ’mlbench’ was built under R version 4.1.3
```

```r
library(dplyr)
library(stats)
library(factoextra)
```

```
## Warning: package ’factoextra’ was built under R version 4.1.3

## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```r
library(kknn)
```

```
## Warning: package ’kknn’ was built under R version 4.1.3

##
## Attaching package: ’kknn’

## The following object is masked from ’package:caret’:
##
##     contr.dummy
```

```r
library(cluster)
library(vegan)
```

```
## Loading required package: permute

## This is vegan 2.5-7

##
## Attaching package: ’vegan’

## The following object is masked from ’package:caret’:
##
##     tolerance
```

```
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

Part a

```
################ Part a -- Data gathering and integration ################

# Name: Heart Attack Analysis & Prediction Dataset
# Category: Health, Classification, Binary Classification
# Link1: https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset
# Link2: http://rstudio-pubs-static.s3.amazonaws.com/24341_184a58191486470cab97acdbbfe78ed5.html

# Data: 303 observations and 14 columns
# Varibales:
# age - age in years
# sex - sex (1 = male; 0 = female)
# cp - chest pain type (0 = typical angina; 1 = atypical angina; 2 = non-anginal pain; 3 = asymptomati
# trtbps - resting blood pressure (in mm Hg on admission to the hospital)
# chol - serum cholesterol in mg/dl
# fbs - fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
# restecg - resting electrocardiographic results (0 = normal; 1 = having ST-T; 2 = hypertrophy)
# thalach - maximum heart rate achieved
# exang - exercise induced angina (1 = yes; 0 = no)
# oldpeak - ST depression induced by exercise relative to rest
# slope - the slope of the peak exercise ST segment (0 = upsloping; 1 = flat; 2 = downsloping)
# caa - number of major vessels (0-3) colored by fluoroscope
# thall - 1 = normal; 2 = fixed defect; 3 = reversible defect
# output - the predicted attribute - diagnosis of heart disease (Value 0 = < 50% diameter narrowing; V

heart = read.csv("heart.csv")
head(heart)
```

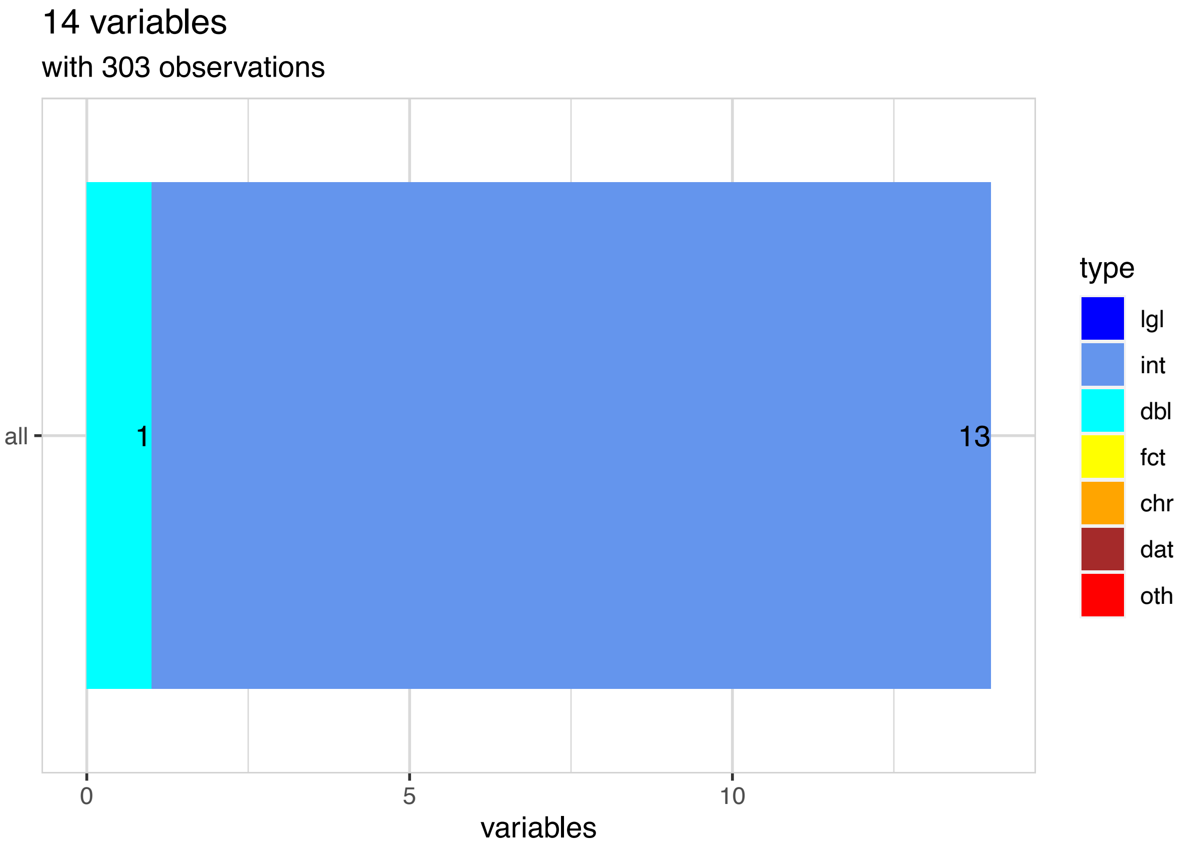
```
##   age sex cp trtbps chol fbs restecg thalachh exng oldpeak slp caa thall output
## 1  63   1  3    145  233   1       0      150    0     2.3   0   0     1      1
## 2  37   1  2    130  250   0       1      187    0     3.5   0   0     2      1
## 3  41   0  1    130  204   0       0      172    0     1.4   2   0     2      1
## 4  56   1  1    120  236   0       1      178    0     0.8   2   0     2      1
## 5  57   0  0    120  354   0       1      163    1     0.6   2   0     2      1
## 6  57   1  0    140  192   0       1      148    0     0.4   1   0     1      1
```

Part b

```
################ Part b -- Data Exploration ################

# b1) summary

# check the data summary
summary(heart)
```

```
##       age             sex               cp            trtbps
##  Min.   :29.00   Min.   :0.0000   Min.   :0.000   Min.   : 94.0
##  1st Qu.:47.50   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:120.0
##  Median :55.00   Median :1.0000   Median :1.000   Median :130.0
##  Mean   :54.37   Mean   :0.6832   Mean   :0.967   Mean   :131.6
##  3rd Qu.:61.00   3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:140.0
##  Max.   :77.00   Max.   :1.0000   Max.   :3.000   Max.   :200.0
##       chol             fbs            restecg          thalachh
##  Min.   :126.0   Min.   :0.0000   Min.   :0.0000   Min.   : 71.0
##  1st Qu.:211.0   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:133.5
##  Median :240.0   Median :0.0000   Median :1.0000   Median :153.0
##  Mean   :246.3   Mean   :0.1485   Mean   :0.5281   Mean   :149.6
##  3rd Qu.:274.5   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:166.0
##  Max.   :564.0   Max.   :1.0000   Max.   :2.0000   Max.   :202.0
##       exng           oldpeak          slp             caa
##  Min.   :0.0000   Min.   :0.00   Min.   :0.000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.00   1st Qu.:1.000   1st Qu.:0.0000
##  Median :0.0000   Median :0.80   Median :1.000   Median :0.0000
##  Mean   :0.3267   Mean   :1.04   Mean   :1.399   Mean   :0.7294
##  3rd Qu.:1.0000   3rd Qu.:1.60   3rd Qu.:2.000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :6.20   Max.   :2.000   Max.   :4.0000
##      thall           output
##  Min.   :0.000   Min.   :0.0000
##  1st Qu.:2.000   1st Qu.:0.0000
##  Median :2.000   Median :1.0000
##  Mean   :2.314   Mean   :0.5446
##  3rd Qu.:3.000   3rd Qu.:1.0000
##  Max.   :3.000   Max.   :1.0000
```

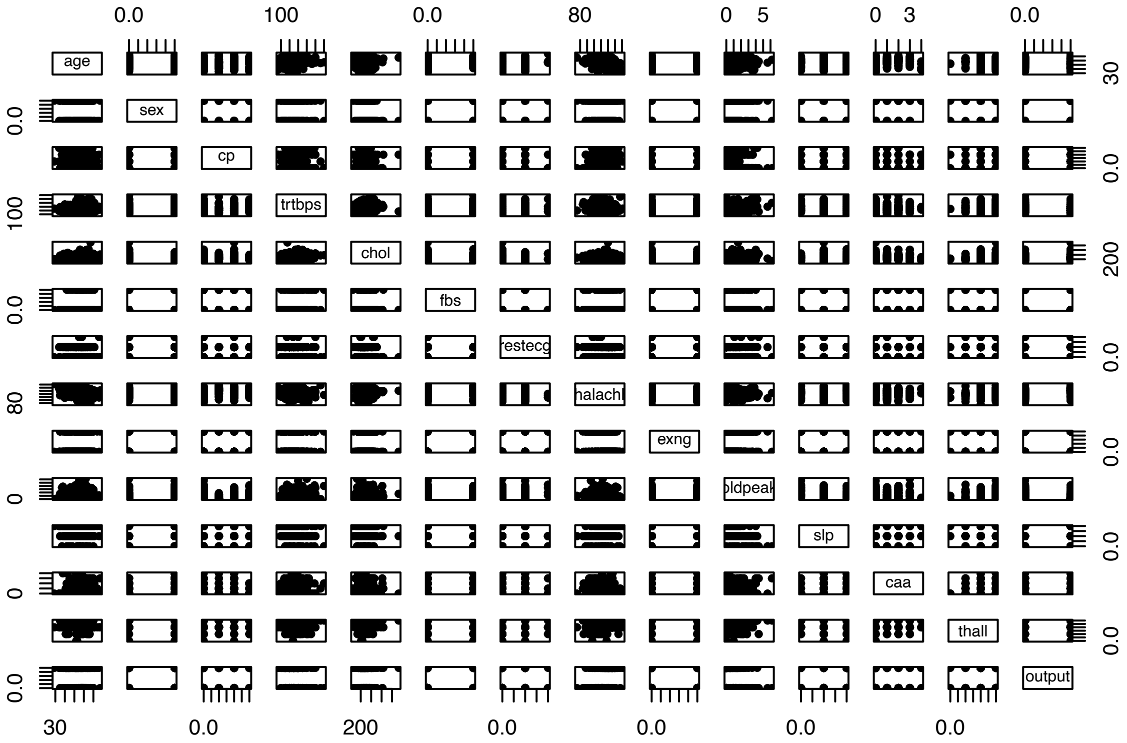
```
# explore dataset
heart %>% explore_tbl()
```



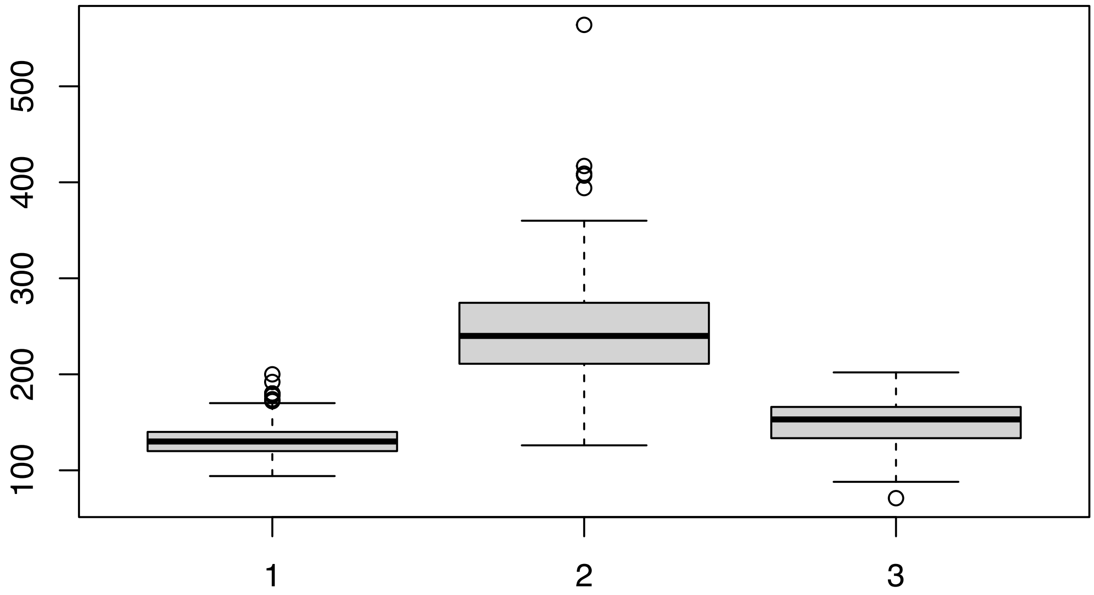
```
heart %>% describe()
```

```
## # A tibble: 14 x 8
##    variable type     na na_pct unique   min   mean   max
##    <chr>    <chr> <int>  <dbl>  <int> <dbl>  <dbl> <dbl>
##  1 age      int       0      0     41    29  54.4   77
##  2 sex      int       0      0      2     0   0.68   1
##  3 cp       int       0      0      4     0   0.97   3
##  4 trtbps   int       0      0     49    94 132.   200
##  5 chol     int       0      0    152   126 246.   564
##  6 fbs      int       0      0      2     0   0.15   1
##  7 restecg  int       0      0      3     0   0.53   2
##  8 thalachh int       0      0     91    71 150.   202
##  9 exng     int       0      0      2     0   0.33   1
## 10 oldpeak  dbl       0      0     40     0   1.04   6.2
## 11 slp      int       0      0      3     0   1.4    2
## 12 caa      int       0      0      5     0   0.73   4
## 13 thall    int       0      0      4     0   2.31   3
## 14 output   int       0      0      2     0   0.54   1
```

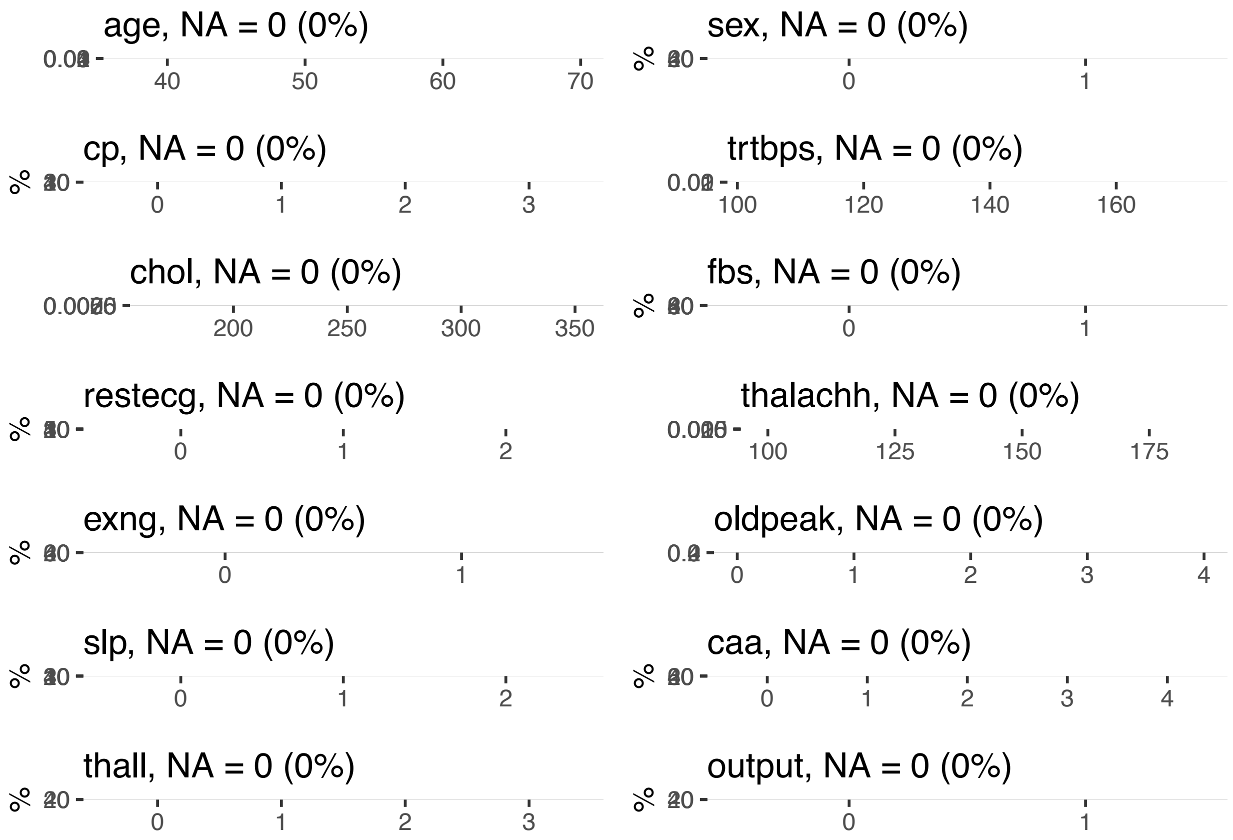
```
# create a scatter plot
pairs(heart[,1:14], pch = 20)
```



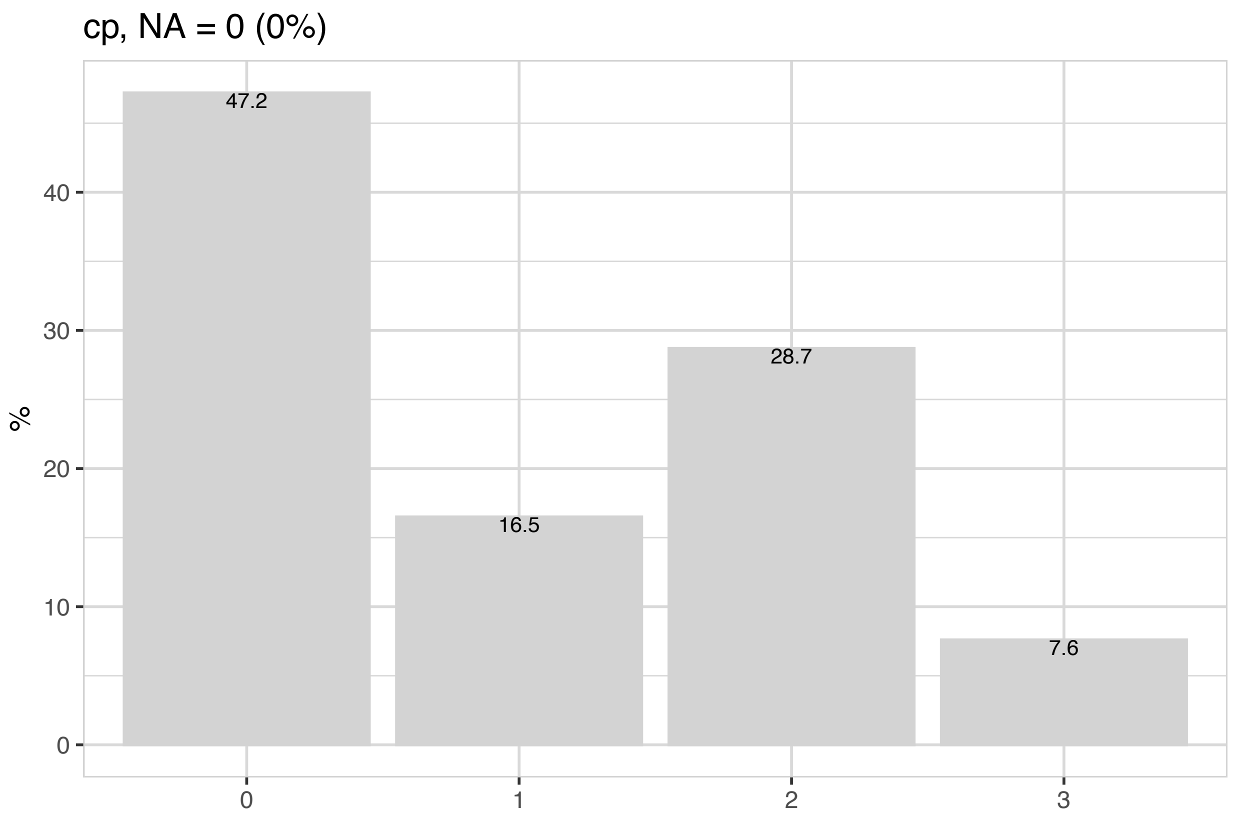
```r
# create box plots
boxplot(heart$trtbps, heart$chol, heart$thalachh)
```



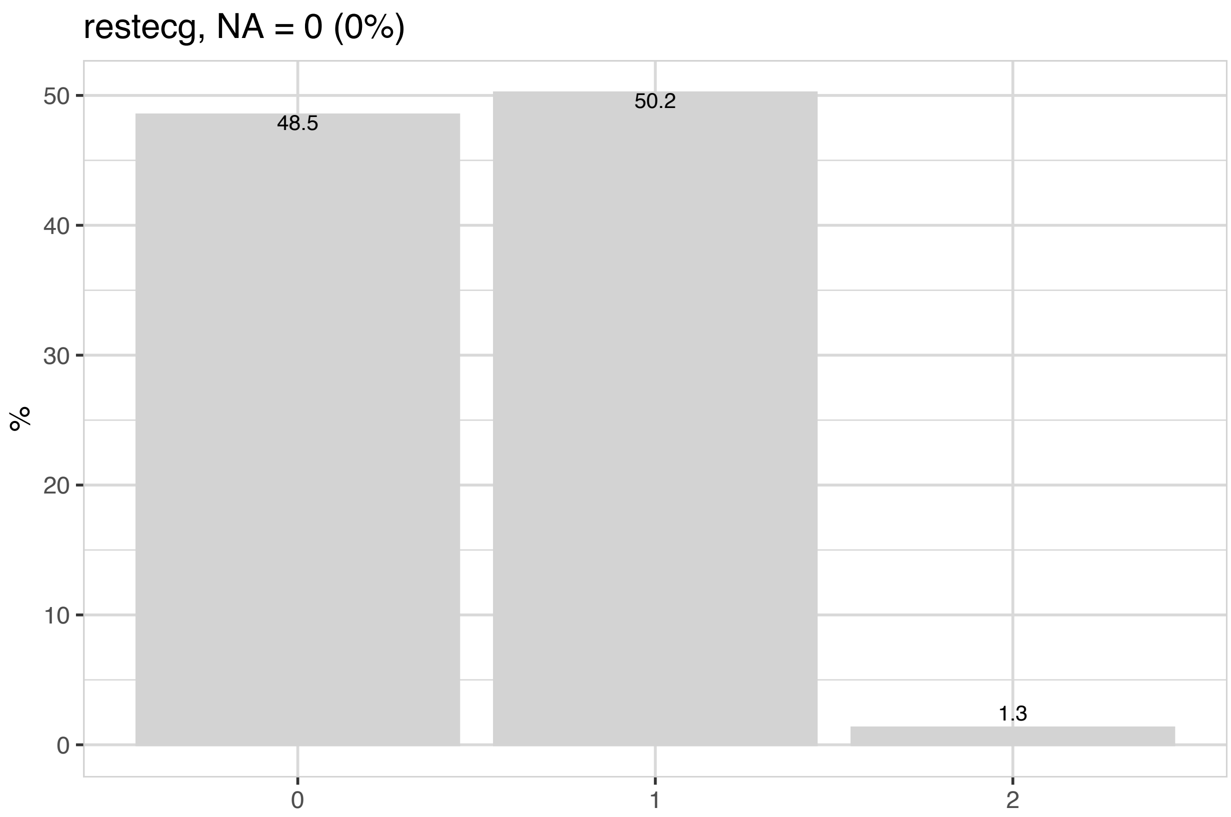
```
# b2) visualization

# explore variables
heart %>% explore_all()
```



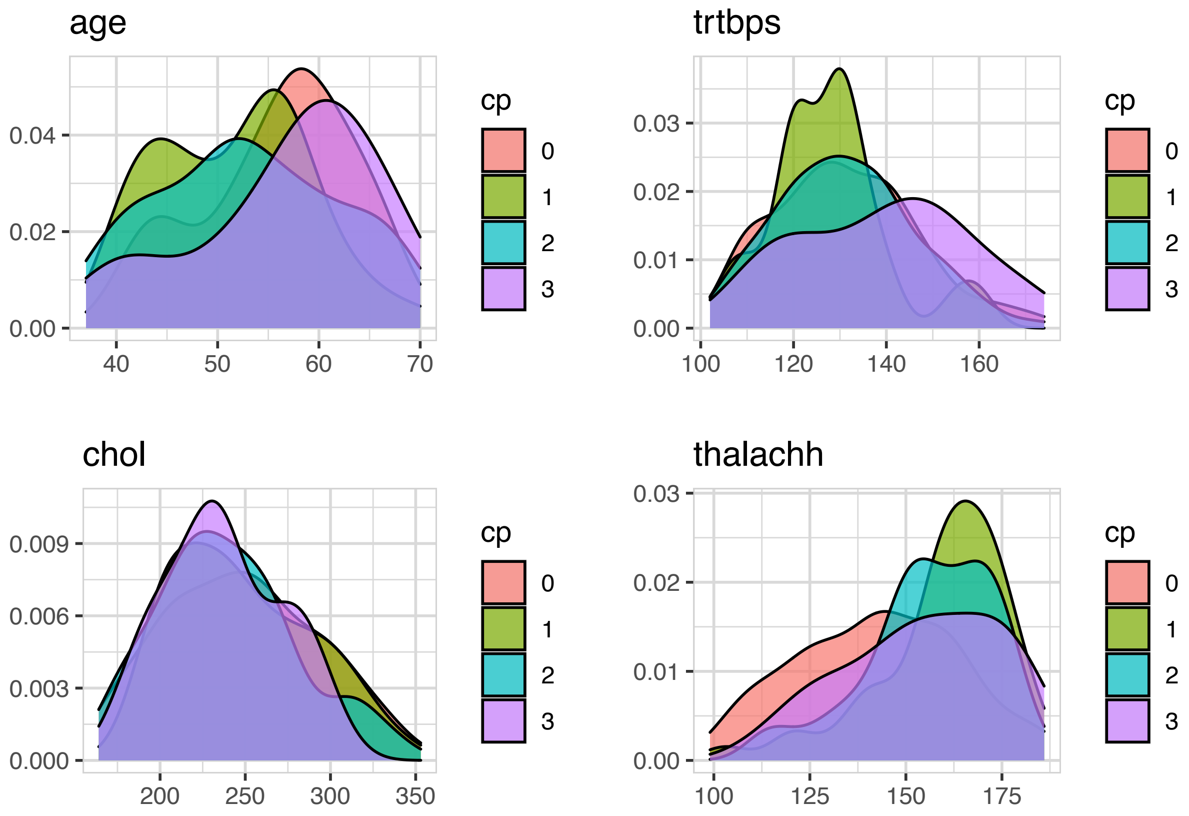
```
# explore chest pain types
heart %>% explore(cp)
```



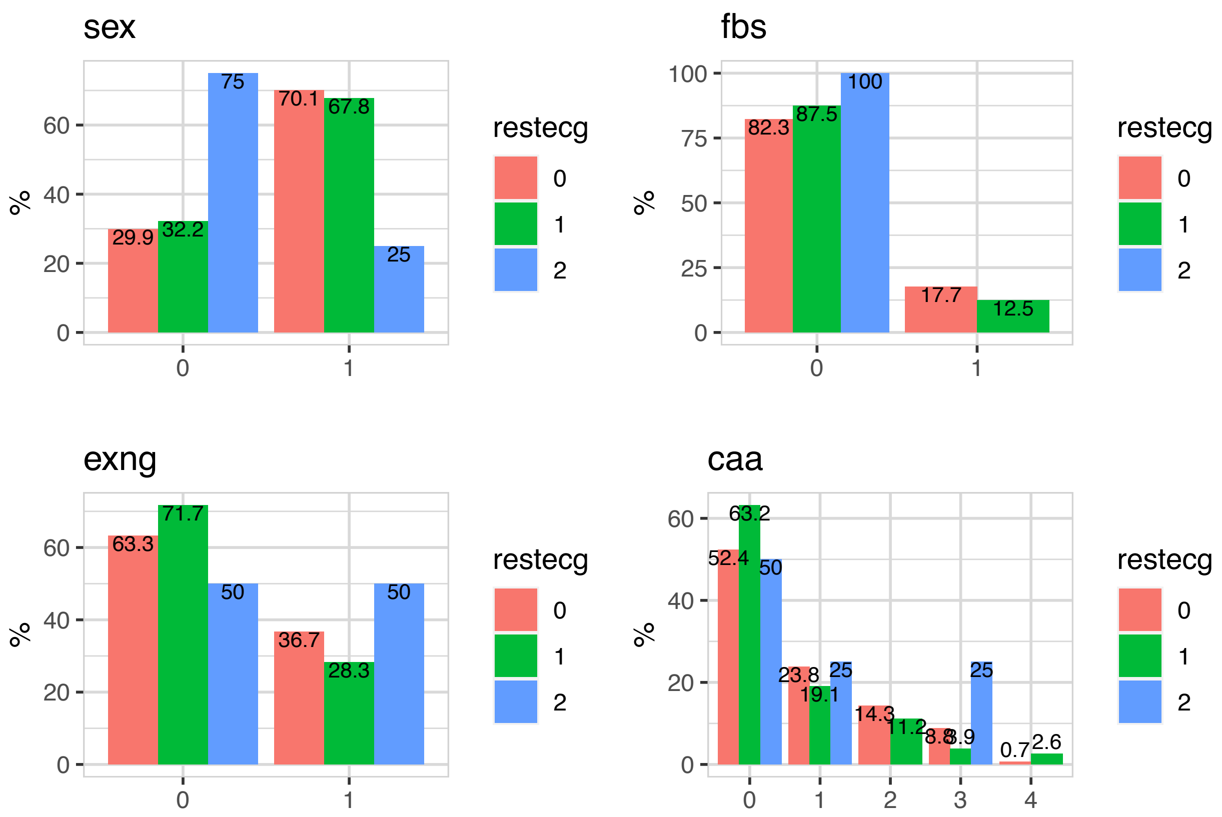
```
# explore resting electroencephalographic results
heart %>% explore(restecg)
```



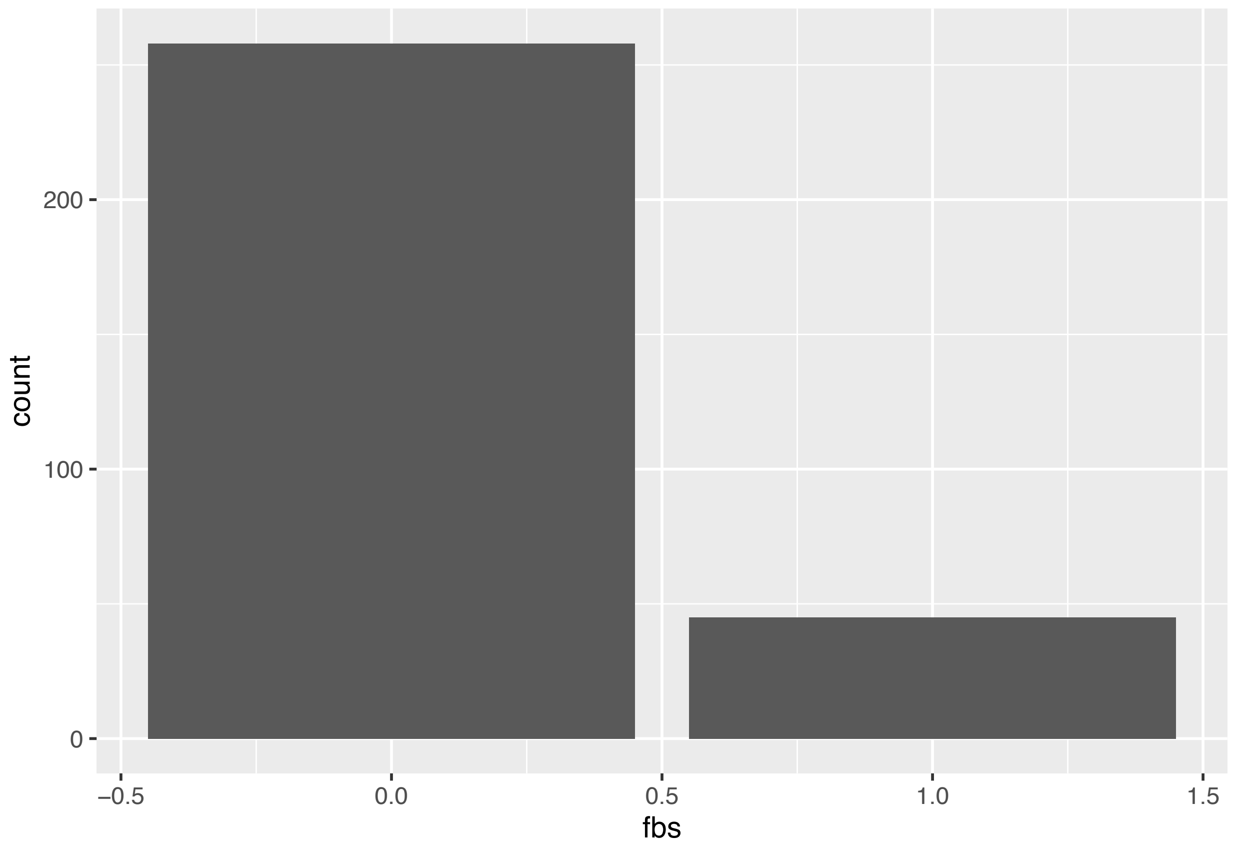
```
# check relation between cp types and four features
heart %>%
  select(cp, age, trtbps, chol, thalachh) %>%
  explore_all(target = cp)
```



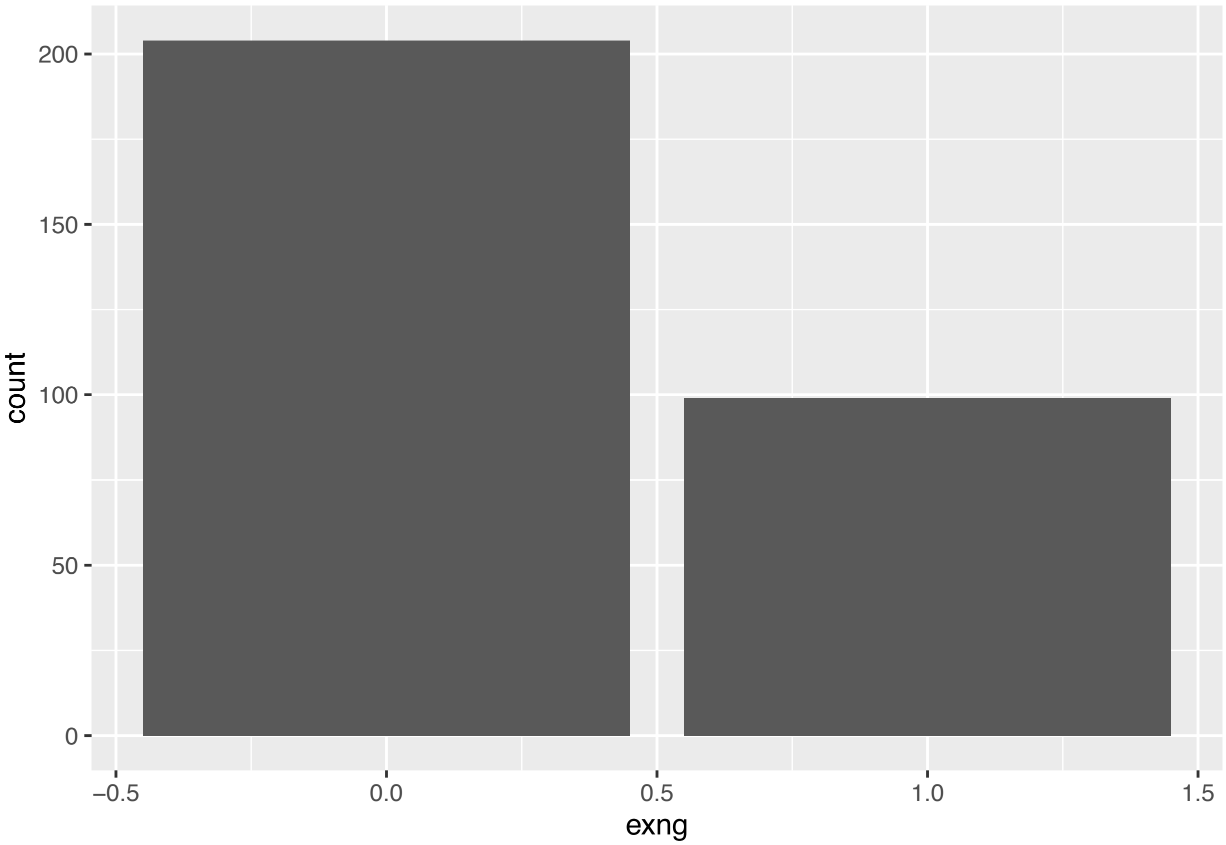
```
# check relation between cp types and four features
heart %>%
  select(restecg, sex, fbs, exng, caa) %>%
  explore_all(target = restecg)
```



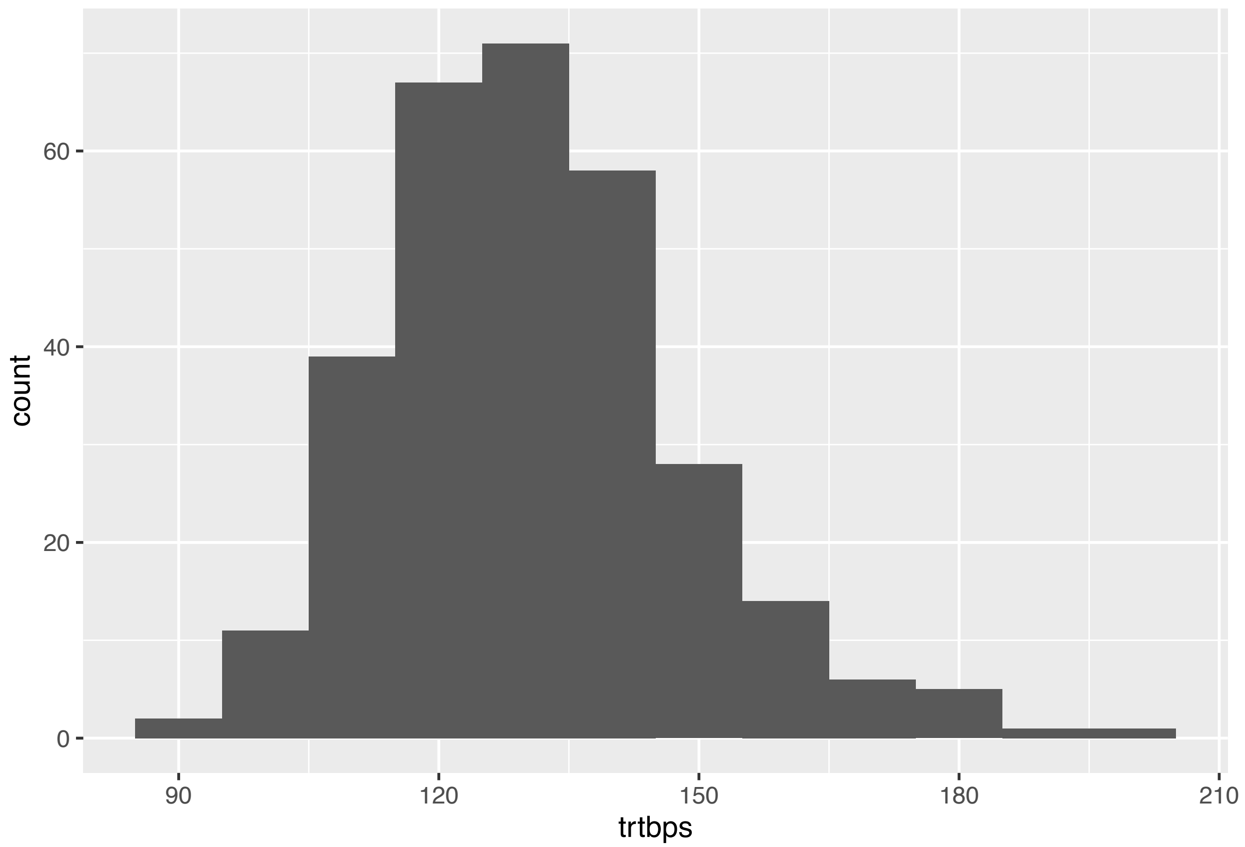
```
# b3) plots

# feature comparison
# geom_bar for binary variable fbs and exng
ggplot(heart, aes(x = fbs)) + geom_bar()
```



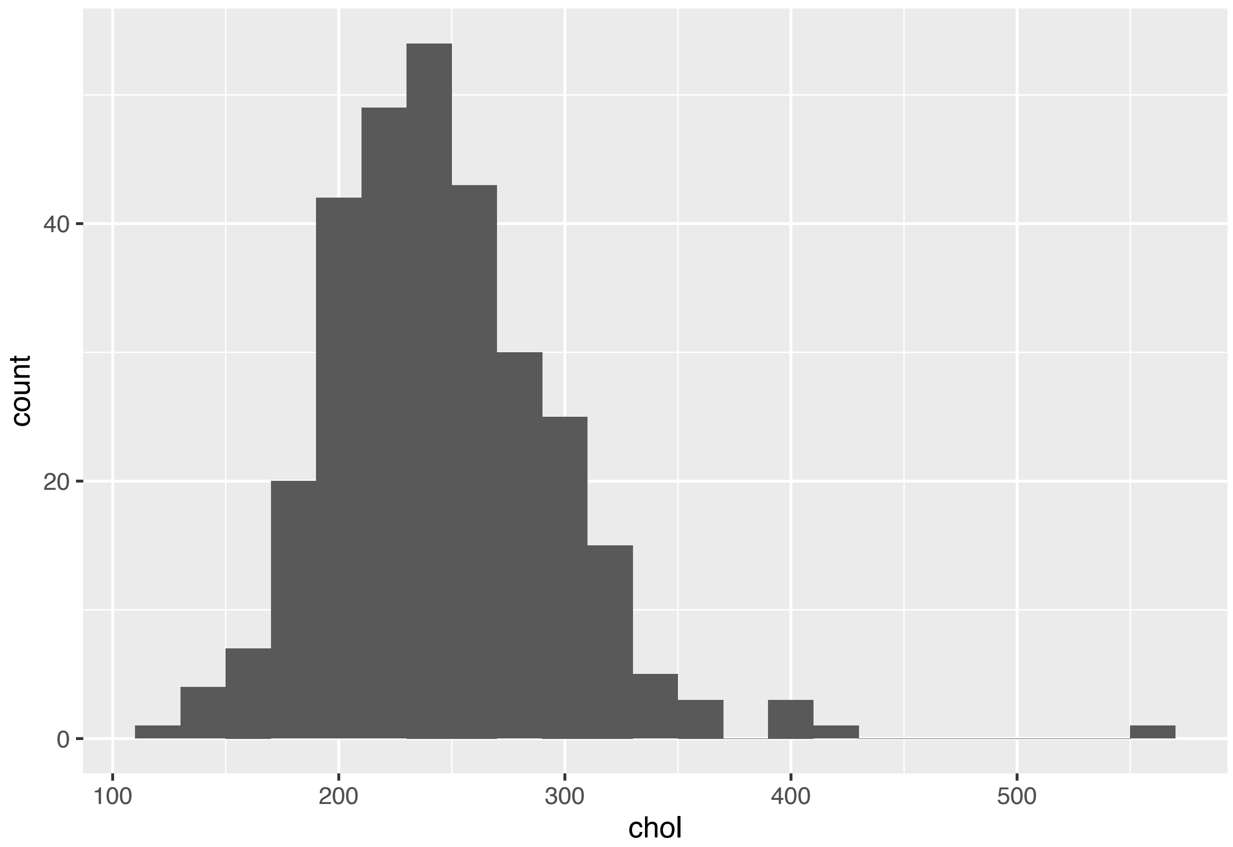
```
ggplot(heart, aes(x = exng)) + geom_bar()
```



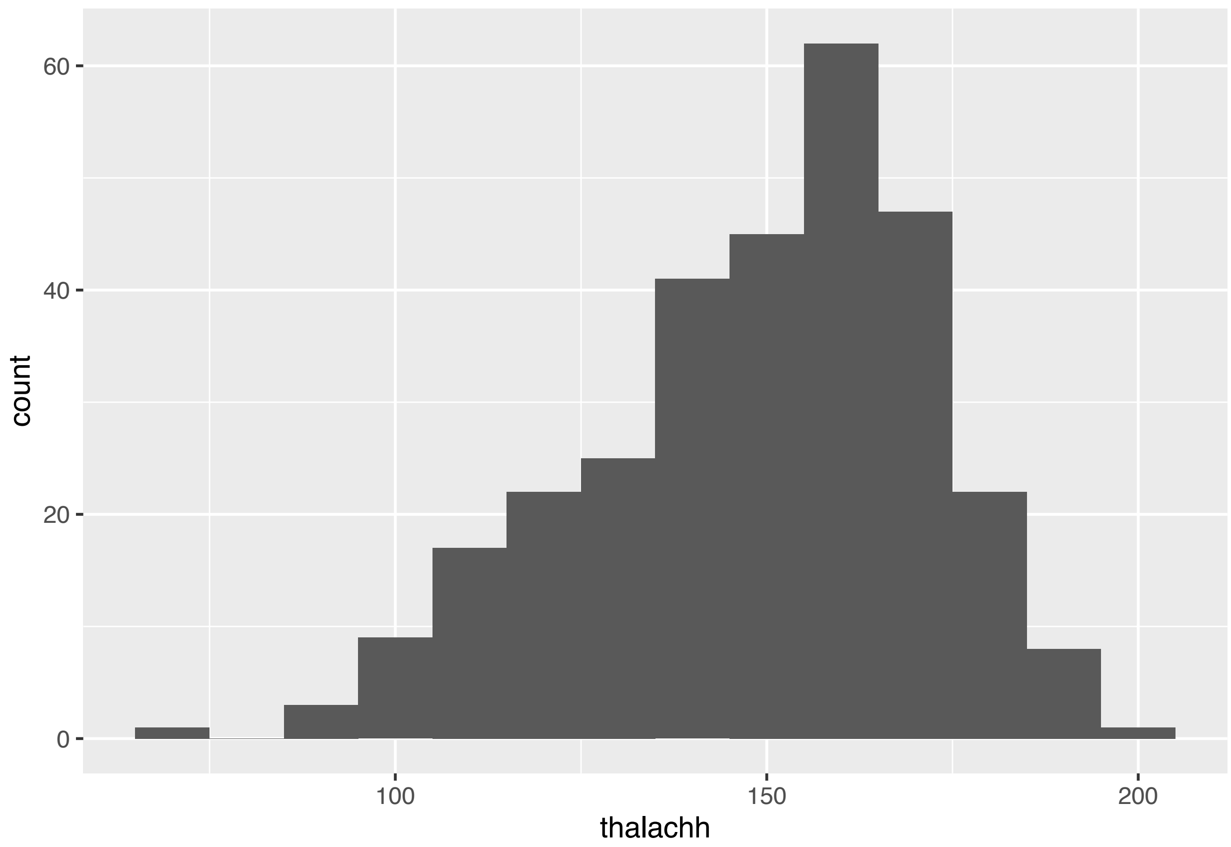
```
# geom_histogram for numerical variable trtbps, chol and thalachh
ggplot(heart, aes(x = trtbps)) + geom_histogram(binwidth = 10)
```



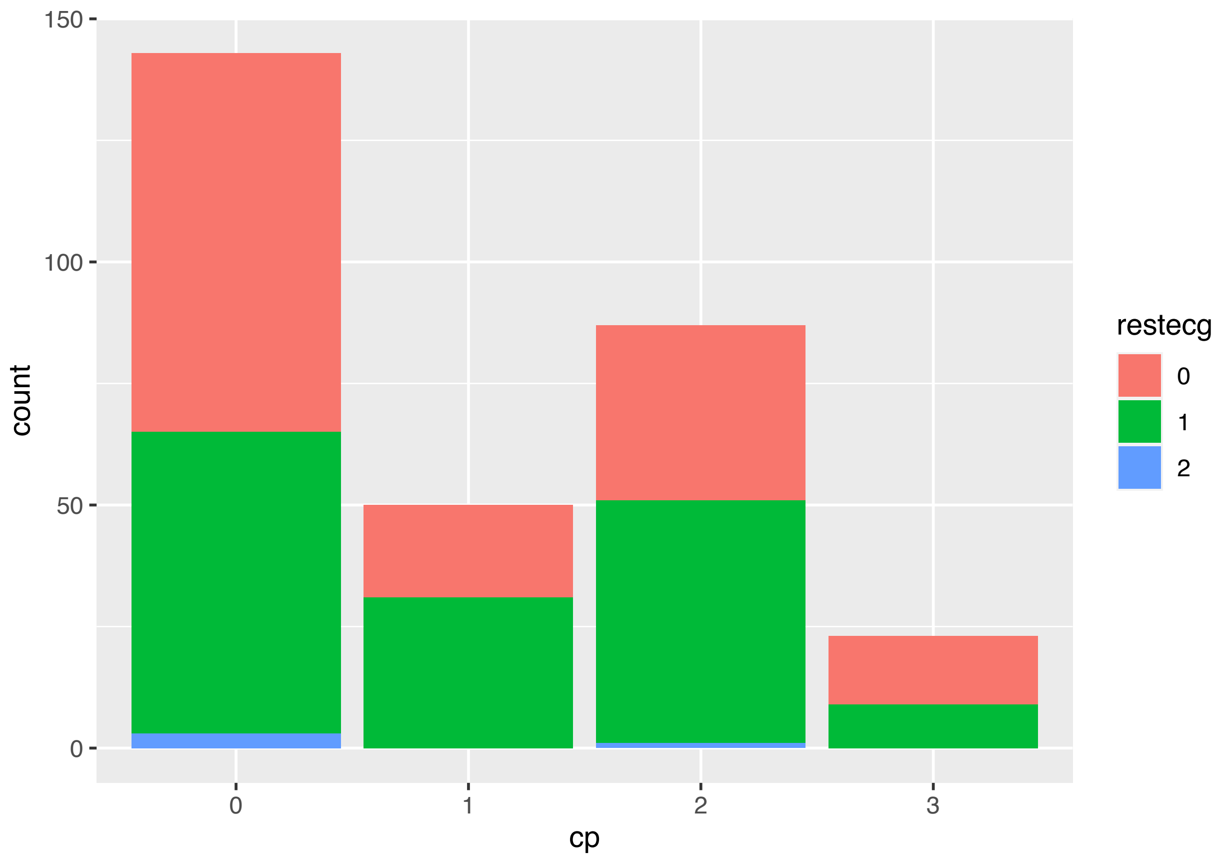
```
ggplot(heart, aes(x = chol)) + geom_histogram(binwidth = 20)
```



```
ggplot(heart, aes(x = thalachh)) + geom_histogram(binwidth = 10)
```



```
# convert cp and restecg to factors
heart$cp <- factor(heart$cp)
heart$restecg <- factor(heart$restecg)

# chest pain type filled with electrocardiograph results
ggplot(heart, aes(x = cp, fill = restecg)) +
  geom_bar(position="stack")
```



Part c

```
################ Part c -- Data Cleaning ################

# c1) detect and clean NAs

# create a new data frame d
d = read.csv("heart.csv")

# check NAs in each column
which(is.na(d))
```

```
## integer(0)
```

```
# check NAs in dataset
sum(is.na(d))
```

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

```
# c2) change variable types

# convert level variables to factors
d$sex <- factor(d$sex)
levels(d$sex) <- c("female", "male")
```

```
d$cp <- factor(d$cp)
levels(d$cp) <- c("typical","atypical","non-anginal","asymptomatic")

d$fbs <- factor(d$fbs)
levels(d$fbs) <- c("false", "true")

d$restecg <- factor(d$restecg)
levels(d$restecg) <- c("normal","stt","hypertrophy")

d$exng <- factor(d$exng)
levels(d$exng) <- c("no","yes")

d$slp <- factor(d$slp)
levels(d$slp) <- c("down","flat","up")

d$caa <- factor(d$caa)

d$thall <- factor(d$thall)
levels(d$thall) <- c("none", "normal","fixed","reversible")

d$output <- factor(d$output)

# c3) remove meaningless columns

# no columns are removed at this phase
# 14 columns are well explained with domain knowledge
```

Part d

```
################ Part d -- Data Pre-processing ################

# d1) normalization on numerical variables

# z-score standardization
preproc1 <- preProcess(d, method=c("center", "scale"))
norm1 <- predict(preproc1, d)
summary(norm1)
```

```
##       age               sex                 cp          trtbps
##  Min.   :-2.79300   female: 96   typical     :143   Min.   :-2.14525
##  1st Qu.:-0.75603   male  :207   atypical    : 50   1st Qu.:-0.66277
##  Median : 0.06977                non-anginal : 87   Median :-0.09259
##  Mean   : 0.00000                asymptomatic: 23   Mean   : 0.00000
##  3rd Qu.: 0.73041                                   3rd Qu.: 0.47760
##  Max.   : 2.49212                                   Max.   : 3.89872
##       chol            fbs             restecg       thalachh        exng
##  Min.   :-2.3203   false:258   normal     :147   Min.   :-3.4336   no :204
##  1st Qu.:-0.6804   true : 45   stt        :152   1st Qu.:-0.7049   yes: 99
##  Median :-0.1209               hypertrophy:  4   Median : 0.1464
##  Mean   : 0.0000                                 Mean   : 0.0000
##  3rd Qu.: 0.5448                                 3rd Qu.: 0.7139
##  Max.   : 6.1303                                 Max.   : 2.2856
##     oldpeak          slp      caa            thall     output
##  Min.   :-0.8954   down: 21   0:175   none      :  2   0:138  
##  1st Qu.:-0.8954   flat:140   1: 65   normal    : 18   1:165  
##  Median :-0.2064   up  :142   2: 38   fixed     :166          
##  Mean   : 0.0000              3: 20   reversible:117          
##  3rd Qu.: 0.4827              4:  5                           
##  Max.   : 4.4445
```

```
# min-max
preproc2 <- preProcess(d, method=c("range"))
norm2 <- predict(preproc2, d)
summary(norm2)
```

```
##       age              sex                    cp           trtbps      
##  Min.   :0.0000   female: 96   typical     :143   Min.   :0.0000  
##  1st Qu.:0.3854   male  :207   atypical    : 50   1st Qu.:0.2453  
##  Median :0.5417                non-anginal : 87   Median :0.3396  
##  Mean   :0.5285                asymptomatic: 23   Mean   :0.3549  
##  3rd Qu.:0.6667                                   3rd Qu.:0.4340  
##  Max.   :1.0000                                   Max.   :1.0000  
##       chol             fbs              restecg       thalachh       exng    
##  Min.   :0.0000   false:258   normal     :147   Min.   :0.0000   no :204  
##  1st Qu.:0.1941   true : 45   stt        :152   1st Qu.:0.4771   yes: 99  
##  Median :0.2603               hypertrophy:  4   Median :0.6260            
##  Mean   :0.2746                                 Mean   :0.6004            
##  3rd Qu.:0.3390                                 3rd Qu.:0.7252            
##  Max.   :1.0000                                 Max.   :1.0000            
##     oldpeak          slp      caa             thall      output 
##  Min.   :0.0000   down: 21   0:175   none      :  2   0:138  
##  1st Qu.:0.0000   flat:140   1: 65   normal    : 18   1:165  
##  Median :0.1290   up  :142   2: 38   fixed     :166          
##  Mean   :0.1677              3: 20   reversible:117          
##  3rd Qu.:0.2581              4:  5                           
##  Max.   :1.0000
```

```
# d2) binning 3 important numerical features: trtbps, chol and thalachh

summary(d)
```

```
##       age             sex                    cp           trtbps     
##  Min.   :29.00   female: 96   typical     :143   Min.   : 94.0  
##  1st Qu.:47.50   male  :207   atypical    : 50   1st Qu.:120.0  
##  Median :55.00                non-anginal : 87   Median :130.0  
##  Mean   :54.37                asymptomatic: 23   Mean   :131.6  
##  3rd Qu.:61.00                                   3rd Qu.:140.0  
##  Max.   :77.00                                   Max.   :200.0  
##       chol            fbs              restecg       thalachh      exng    
##  Min.   :126.0   false:258   normal     :147   Min.   : 71.0   no :204  
##  1st Qu.:211.0   true : 45   stt        :152   1st Qu.:133.5   yes: 99  
##  Median :240.0               hypertrophy:  4   Median :153.0            
##  Mean   :246.3                                 Mean   :149.6            
##  3rd Qu.:274.5                                 3rd Qu.:166.0            
##  Max.   :564.0                                 Max.   :202.0            
##     oldpeak         slp      caa             thall      output 
##  Min.   :0.00   down: 21   0:175   none      :  2   0:138
##  1st Qu.:0.00   flat:140   1: 65   normal    : 18   1:165
##  Median :0.80   up  :142   2: 38   fixed     :166
##  Mean   :1.04              3: 20   reversible:117
##  3rd Qu.:1.60              4:  5
##  Max.   :6.20
```

```r
df <- d %>%
  mutate(trtbps_bin = cut(trtbps,
                        breaks=c(90, 120, 140, 200),
                        labels=c("low","medium","high")))

df <- df %>%
  mutate(chol_bin = cut(chol,
                      breaks=c(120, 220, 260, 580),
                      labels=c("low","medium","high")))

df <- df %>%
  mutate(thalachh_bin = cut(thalachh,
                        breaks=3,
                        labels=c("low","medium","high")))

# d3) smoothing binned features

# smoothing trtbps_bin and replace values
low_trtbps <- df %>%
  filter(trtbps_bin == "low") %>%
  mutate(trtbps = mean(trtbps))
medium_trtbps <- df %>%
  filter(trtbps_bin == "medium") %>%
  mutate(trtbps = mean(trtbps))
high_trtbps <- df %>%
  filter(trtbps_bin == "high") %>%
  mutate(trtbps = mean(trtbps))
# Tidyverse to combine these separate sets
new_trtbps <- bind_rows(list(low_trtbps, medium_trtbps, high_trtbps))

# smoothing chol_bin and replace values
low_chol <- df %>%
  filter(chol_bin == "low") %>%
  mutate(chol = mean(chol))
medium_chol <- df %>%
  filter(chol_bin == "medium") %>%
  mutate(chol = mean(chol))
high_chol <- df %>%
  filter(chol_bin == "high") %>%
  mutate(chol = mean(chol))
# Tidyverse to combine these separate sets
new_chol <- bind_rows(list(low_chol, medium_chol, high_chol))

# smoothing thalachh_bin and replace values
low_thalachh <- df %>%
  filter(thalachh_bin == "low") %>%
  mutate(thalachh = mean(thalachh))
```

```r
medium_thalachh <- df %>%
  filter(thalachh_bin == "medium") %>%
  mutate(thalachh = mean(thalachh))
high_thalachh <- df %>%
  filter(thalachh_bin == "high") %>%
  mutate(thalachh = mean(thalachh))
# Tidyverse to combine these separate sets
new_thalachh <- bind_rows(list(low_thalachh, medium_thalachh, high_thalachh))


# d4) replace smoothed columns to form the final pre-processed data frame

df_heart <- df
# replace trtbps column
df_heart$trtbps <- new_trtbps$trtbps
# replace chol column
df_heart$chol <- new_chol$chol
# replace thalachh column
df_heart$thalachh <- new_thalachh$thalachh
# present updated df_heart
df_heart
```

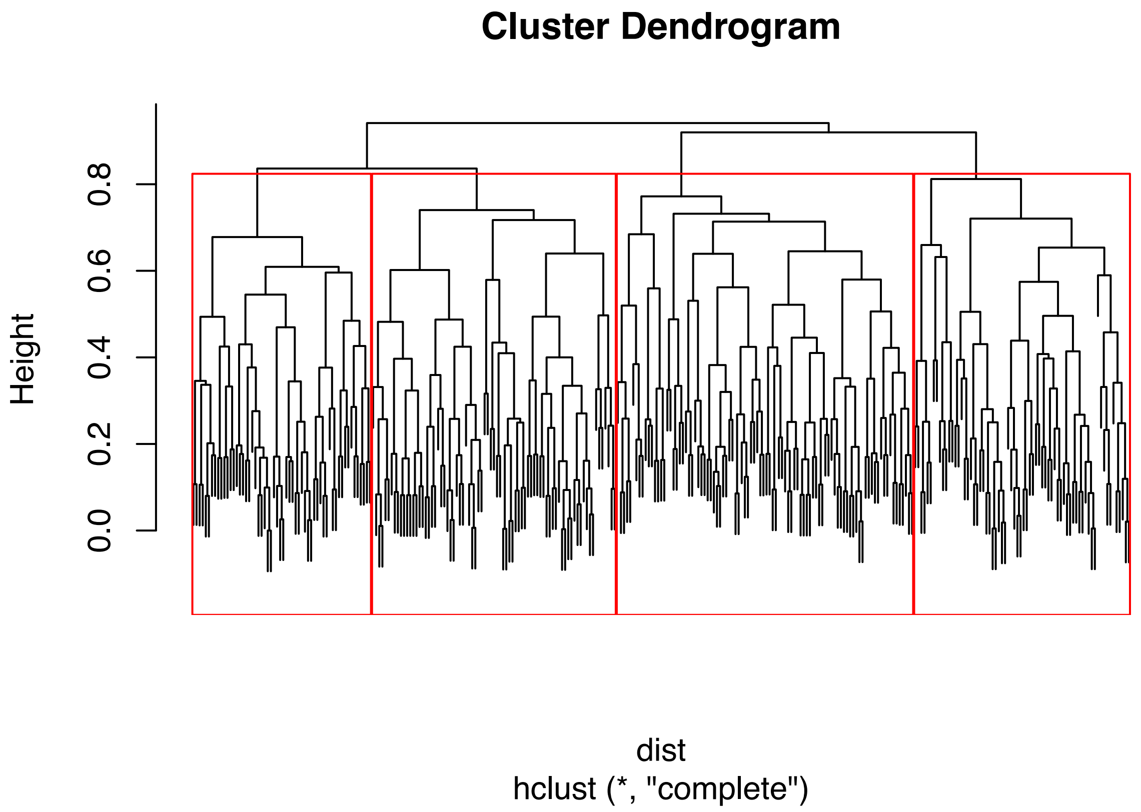
```
##     age    sex           cp   trtbps     chol   fbs     restecg thalachh exng
## 1    63   male asymptomatic 113.4639 194.6061  true      normal 103.8889   no
## 2    37   male  non-anginal 113.4639 194.6061 false         stt 103.8889   no
## 3    41 female     atypical 113.4639 194.6061 false      normal 103.8889   no
## 4    56   male     atypical 113.4639 194.6061 false         stt 103.8889   no
## 5    57 female      typical 113.4639 194.6061 false         stt 103.8889  yes
## 6    57   male      typical 113.4639 194.6061 false         stt 103.8889   no
## 7    56 female     atypical 113.4639 194.6061 false      normal 103.8889   no
## 8    44   male     atypical 113.4639 194.6061 false         stt 103.8889   no
## 9    52   male  non-anginal 113.4639 194.6061  true         stt 103.8889   no
## 10   57   male  non-anginal 113.4639 194.6061 false         stt 103.8889   no
## 11   54   male      typical 113.4639 194.6061 false         stt 103.8889   no
## 12   48 female  non-anginal 113.4639 194.6061 false         stt 103.8889   no
## 13   49   male     atypical 113.4639 194.6061 false         stt 103.8889   no
## 14   64   male asymptomatic 113.4639 194.6061 false      normal 103.8889  yes
## 15   58 female asymptomatic 113.4639 194.6061  true      normal 103.8889   no
## 16   50 female  non-anginal 113.4639 194.6061 false         stt 103.8889   no
## 17   58 female  non-anginal 113.4639 194.6061 false         stt 103.8889   no
## 18   66 female asymptomatic 113.4639 194.6061 false         stt 103.8889   no
## 19   43   male      typical 113.4639 194.6061 false         stt 103.8889   no
## 20   69 female asymptomatic 113.4639 194.6061 false         stt 103.8889   no
## 21   59   male      typical 113.4639 194.6061 false         stt 103.8889   no
## 22   44   male  non-anginal 113.4639 194.6061 false         stt 103.8889  yes
## 23   42   male      typical 113.4639 194.6061 false         stt 103.8889   no
## 24   61   male  non-anginal 113.4639 194.6061  true         stt 103.8889  yes
## 25   40   male asymptomatic 113.4639 194.6061 false         stt 103.8889  yes
## 26   71 female     atypical 113.4639 194.6061 false         stt 103.8889   no
## 27   59   male  non-anginal 113.4639 194.6061  true         stt 103.8889   no
## 28   51   male  non-anginal 113.4639 194.6061 false         stt 140.7582   no
## 29   65 female  non-anginal 113.4639 194.6061  true      normal 140.7582   no
## 30   53   male  non-anginal 113.4639 194.6061  true      normal 140.7582   no
## 31   41 female     atypical 113.4639 194.6061 false         stt 140.7582   no
## 32   65   male      typical 113.4639 194.6061 false         stt 140.7582   no
## 33   44   male     atypical 113.4639 194.6061 false      normal 140.7582   no
## 34   54   male  non-anginal 113.4639 194.6061 false      normal 140.7582   no
## 35   51   male asymptomatic 113.4639 194.6061 false      normal 140.7582  yes
## 36   46 female  non-anginal 113.4639 194.6061 false      normal 140.7582  yes
## 37   54 female  non-anginal 113.4639 194.6061  true         stt 140.7582   no
## 38   54   male  non-anginal 113.4639 194.6061 false      normal 140.7582   no
## 39   65 female  non-anginal 113.4639 194.6061 false         stt 140.7582   no
## 40   65 female  non-anginal 113.4639 194.6061 false      normal 140.7582   no
## 41   51 female  non-anginal 113.4639 194.6061 false      normal 140.7582   no
## 42   48   male     atypical 113.4639 194.6061 false      normal 140.7582   no
## 43   45   male      typical 113.4639 194.6061 false      normal 140.7582  yes
## 44   53 female      typical 113.4639 194.6061 false      normal 140.7582   no
## 45   39   male  non-anginal 113.4639 194.6061 false      normal 140.7582   no
## 46   52   male     atypical 113.4639 194.6061 false         stt 140.7582   no
## 47   44   male  non-anginal 113.4639 194.6061 false      normal 140.7582   no
## 48   47   male  non-anginal 113.4639 194.6061 false      normal 140.7582   no
## 49   53 female  non-anginal 113.4639 194.6061 false      normal 140.7582   no
## 50   53 female      typical 113.4639 194.6061 false      normal 140.7582   no
## 51   51 female  non-anginal 113.4639 194.6061 false      normal 140.7582   no
## 52   66   male      typical 113.4639 194.6061 false      normal 140.7582   no
## 53   62   male  non-anginal 113.4639 194.6061 false         stt 140.7582   no
## 54   44 female  non-anginal 113.4639 194.6061 false         stt 140.7582   no
## 55   63 female  non-anginal 113.4639 194.6061 false      normal 140.7582   no
## 56   52   male     atypical 113.4639 194.6061 false         stt 140.7582   no
## 57   48   male      typical 113.4639 194.6061 false      normal 140.7582   no
## 58   45   male      typical 113.4639 194.6061 false      normal 140.7582   no
## 59   34   male asymptomatic 113.4639 194.6061 false      normal 140.7582   no
## 60   57 female      typical 113.4639 194.6061 false      normal 140.7582   no
## 61   71 female  non-anginal 113.4639 194.6061  true      normal 140.7582   no
## 62   54   male     atypical 113.4639 194.6061 false         stt 140.7582   no
## 63   52   male asymptomatic 113.4639 194.6061 false      normal 140.7582   no
## 64   41   male     atypical 113.4639 194.6061 false         stt 140.7582   no
## 65   58   male  non-anginal 113.4639 194.6061  true      normal 140.7582   no
## 66   35 female      typical 113.4639 194.6061 false         stt 140.7582   no
## 67   51   male  non-anginal 113.4639 194.6061 false         stt 140.7582  yes
## 68   45 female     atypical 113.4639 194.6061 false      normal 140.7582   no
## 69   44   male     atypical 113.4639 194.6061 false         stt 140.7582   no
## 70   62 female      typical 113.4639 194.6061 false         stt 140.7582   no
## 71   54   male  non-anginal 113.4639 194.6061 false      normal 140.7582   no
## 72   51   male  non-anginal 113.4639 194.6061 false         stt 140.7582  yes
## 73   29   male     atypical 113.4639 194.6061 false      normal 140.7582   no
## 74   51   male      typical 113.4639 194.6061 false      normal 140.7582  yes
## 75   43 female  non-anginal 113.4639 194.6061 false         stt 140.7582   no
## 76   55 female     atypical 113.4639 194.6061 false      normal 140.7582   no
## 77   51   male  non-anginal 113.4639 194.6061  true      normal 140.7582   no
## 78   59   male     atypical 113.4639 194.6061 false         stt 140.7582  yes
## 79   52   male     atypical 113.4639 194.6061  true         stt 140.7582   no
## 80   58   male  non-anginal 113.4639 194.6061 false      normal 140.7582  yes
## 81   41   male  non-anginal 113.4639 194.6061 false         stt 140.7582   no
## 82   45   male     atypical 113.4639 194.6061 false      normal 140.7582   no
## 83   60 female  non-anginal 113.4639 194.6061 false         stt 140.7582   no
## 84   52   male asymptomatic 113.4639 194.6061  true         stt 140.7582   no
## 85   42 female      typical 113.4639 194.6061 false      normal 140.7582   no
## 86   67 female  non-anginal 113.4639 194.6061 false      normal 140.7582   no
## 87   68   male  non-anginal 113.4639 194.6061 false         stt 140.7582   no
## 88   46   male     atypical 113.4639 194.6061  true         stt 140.7582   no
## 89   54 female  non-anginal 113.4639 194.6061 false         stt 140.7582   no
## 90   58 female      typical 113.4639 194.6061 false      normal 140.7582   no
## 91   48   male  non-anginal 113.4639 194.6061  true         stt 140.7582   no
## 92   57   male      typical 113.4639 194.6061 false         stt 140.7582  yes
## 93   52   male  non-anginal 113.4639 194.6061 false         stt 140.7582   no
## 94   54 female     atypical 113.4639 194.6061  true      normal 140.7582  yes
## 95   45 female     atypical 113.4639 194.6061 false         stt 140.7582   no
## 96   53   male      typical 113.4639 194.6061 false      normal 140.7582  yes
## 97   62 female      typical 113.4639 194.6061 false      normal 140.7582   no
## 98   52   male      typical 132.4184 194.6061  true         stt 140.7582   no
## 99   43   male  non-anginal 132.4184 194.6061 false         stt 140.7582   no
## 100  53   male  non-anginal 132.4184 239.8673  true      normal 140.7582   no
## 101  42   male asymptomatic 132.4184 239.8673 false      normal 140.7582   no
## 102  59   male asymptomatic 132.4184 239.8673 false      normal 140.7582   no
## 103  63 female     atypical 132.4184 239.8673 false         stt 140.7582   no
## 104  42   male  non-anginal 132.4184 239.8673  true         stt 140.7582   no
## 105  50   male  non-anginal 132.4184 239.8673 false         stt 140.7582   no
## 106  68 female  non-anginal 132.4184 239.8673 false      normal 140.7582   no
## 107  69   male asymptomatic 132.4184 239.8673  true      normal 140.7582   no
## 108  45 female      typical 132.4184 239.8673 false      normal 140.7582  yes
## 109  50 female     atypical 132.4184 239.8673 false         stt 140.7582   no
## 110  50 female      typical 132.4184 239.8673 false      normal 140.7582   no
## 111  64 female      typical 132.4184 239.8673 false         stt 140.7582  yes
## 112  57   male  non-anginal 132.4184 239.8673  true         stt 140.7582   no
## 113  64 female  non-anginal 132.4184 239.8673 false         stt 140.7582   no
## 114  43   male      typical 132.4184 239.8673 false         stt 140.7582   no
## 115  55   male     atypical 132.4184 239.8673 false         stt 140.7582   no
## 116  37 female  non-anginal 132.4184 239.8673 false         stt 140.7582   no
## 117  41   male  non-anginal 132.4184 239.8673 false      normal 140.7582   no
## 118  56   male asymptomatic 132.4184 239.8673 false      normal 140.7582   no
## 119  46 female     atypical 132.4184 239.8673 false         stt 140.7582   no
## 120  46 female      typical 132.4184 239.8673 false      normal 140.7582  yes
## 121  64 female      typical 132.4184 239.8673 false         stt 140.7582   no
## 122  59   male      typical 132.4184 239.8673 false      normal 140.7582   no
## 123  41 female  non-anginal 132.4184 239.8673 false      normal 140.7582  yes
## 124  54 female  non-anginal 132.4184 239.8673 false      normal 140.7582   no
## 125  39 female  non-anginal 132.4184 239.8673 false         stt 140.7582   no
## 126  34 female     atypical 132.4184 239.8673 false         stt 140.7582   no
## 127  47   male      typical 132.4184 239.8673 false         stt 140.7582   no
## 128  67 female  non-anginal 132.4184 239.8673 false         stt 140.7582   no
## 129  52 female  non-anginal 132.4184 239.8673 false      normal 140.7582   no
## 130  74 female     atypical 132.4184 239.8673 false      normal 140.7582  yes
## 131  54 female  non-anginal 132.4184 239.8673 false         stt 140.7582   no
## 132  49 female     atypical 132.4184 239.8673 false         stt 140.7582   no
## 133  42   male     atypical 132.4184 239.8673 false         stt 140.7582   no
## 134  41   male     atypical 132.4184 239.8673 false         stt 140.7582   no
## 135  41 female     atypical 132.4184 239.8673 false         stt 140.7582   no
## 136  49 female      typical 132.4184 239.8673 false         stt 140.7582   no
## 137  60 female  non-anginal 132.4184 239.8673  true         stt 140.7582   no
## 138  62   male     atypical 132.4184 239.8673  true      normal 140.7582   no
## 139  57   male      typical 132.4184 239.8673 false         stt 140.7582  yes
## 140  64   male      typical 132.4184 239.8673 false         stt 140.7582  yes
## 141  51 female  non-anginal 132.4184 239.8673 false      normal 140.7582   no
## 142  43   male      typical 132.4184 239.8673 false         stt 140.7582   no
## 143  42 female  non-anginal 132.4184 239.8673 false         stt 140.7582   no
## 144  67 female      typical 132.4184 239.8673 false         stt 140.7582   no
## 145  76 female  non-anginal 132.4184 239.8673 false hypertrophy 140.7582   no
## 146  70   male     atypical 132.4184 239.8673 false      normal 140.7582   no
## 147  44 female  non-anginal 132.4184 239.8673 false         stt 140.7582   no
## 148  60 female asymptomatic 132.4184 239.8673 false         stt 140.7582   no
## 149  44   male  non-anginal 132.4184 239.8673 false         stt 140.7582   no
## 150  42   male  non-anginal 132.4184 239.8673 false         stt 140.7582   no
## 151  66   male      typical 132.4184 239.8673 false      normal 140.7582   no
## 152  71 female      typical 132.4184 239.8673 false         stt 140.7582   no
## 153  64   male asymptomatic 132.4184 239.8673 false      normal 140.7582   no
## 154  66 female  non-anginal 132.4184 239.8673 false      normal 140.7582   no
## 155  39 female  non-anginal 132.4184 239.8673 false         stt 140.7582   no
## 156  58 female      typical 132.4184 239.8673 false         stt 140.7582   no
## 157  47   male  non-anginal 132.4184 239.8673 false         stt 140.7582   no
## 158  35   male     atypical 132.4184 239.8673 false         stt 140.7582   no
## 159  58   male     atypical 132.4184 239.8673 false         stt 140.7582   no
## 160  56   male     atypical 132.4184 239.8673 false      normal 140.7582   no
## 161  56   male     atypical 132.4184 239.8673 false         stt 140.7582   no
## 162  55 female     atypical 132.4184 239.8673 false         stt 140.7582   no
## 163  41   male     atypical 132.4184 239.8673 false         stt 140.7582   no
## 164  38   male  non-anginal 132.4184 239.8673 false         stt 140.7582   no
## 165  38   male  non-anginal 132.4184 239.8673 false         stt 140.7582   no
## 166  67   male      typical 132.4184 239.8673 false      normal 140.7582  yes
## 167  67   male      typical 132.4184 239.8673 false      normal 140.7582  yes
## 168  62 female      typical 132.4184 239.8673 false      normal 140.7582   no
## 169  63   male      typical 132.4184 239.8673 false      normal 140.7582   no
## 170  53   male      typical 132.4184 239.8673  true      normal 140.7582  yes
## 171  56   male  non-anginal 132.4184 239.8673  true      normal 140.7582  yes
## 172  48   male     atypical 132.4184 239.8673 false         stt 140.7582   no
## 173  58   male     atypical 132.4184 239.8673 false      normal 140.7582   no
## 174  58   male  non-anginal 132.4184 239.8673 false      normal 140.7582   no
## 175  60   male      typical 132.4184 239.8673 false      normal 140.7582  yes
## 176  40   male      typical 132.4184 239.8673 false      normal 140.7582  yes
## 177  60   male      typical 132.4184 239.8673  true         stt 140.7582  yes
## 178  64   male  non-anginal 132.4184 239.8673 false         stt 140.7582   no
## 179  43   male      typical 132.4184 239.8673 false      normal 140.7582  yes
## 180  57   male      typical 132.4184 239.8673 false      normal 140.7582  yes
## 181  55   male      typical 132.4184 239.8673 false         stt 170.7480  yes
## 182  65 female      typical 132.4184 239.8673 false      normal 170.7480   no
## 183  61 female      typical 132.4184 239.8673 false      normal 170.7480   no
## 184  58   male  non-anginal 132.4184 239.8673 false      normal 170.7480   no
## 185  50   male      typical 132.4184 239.8673 false      normal 170.7480   no
## 186  44   male      typical 132.4184 239.8673 false      normal 170.7480   no
## 187  60   male      typical 132.4184 239.8673 false         stt 170.7480  yes
## 188  54   male      typical 132.4184 239.8673 false      normal 170.7480  yes
## 189  50   male  non-anginal 132.4184 239.8673 false         stt 170.7480   no
## 190  41   male      typical 132.4184 239.8673 false      normal 170.7480   no
## 191  51 female      typical 132.4184 239.8673 false         stt 170.7480  yes
## 192  58   male      typical 132.4184 239.8673 false      normal 170.7480  yes
## 193  54   male      typical 132.4184 239.8673 false         stt 170.7480   no
## 194  60   male      typical 132.4184 239.8673 false      normal 170.7480  yes
## 195  60   male  non-anginal 132.4184 239.8673 false      normal 170.7480   no
## 196  59   male      typical 132.4184 239.8673 false      normal 170.7480  yes
## 197  46   male  non-anginal 132.4184 239.8673 false         stt 170.7480   no
## 198  67   male      typical 132.4184 300.4245  true         stt 170.7480   no
## 199  62   male      typical 132.4184 300.4245 false         stt 170.7480  yes
## 200  65   male      typical 132.4184 300.4245 false      normal 170.7480   no
## 201  44   male      typical 132.4184 300.4245 false      normal 170.7480   no
## 202  60   male      typical 132.4184 300.4245 false      normal 170.7480  yes
## 203  58   male      typical 132.4184 300.4245 false      normal 170.7480  yes
## 204  68   male  non-anginal 132.4184 300.4245  true      normal 170.7480  yes
## 205  62 female      typical 132.4184 300.4245 false      normal 170.7480   no
## 206  52   male      typical 132.4184 300.4245 false         stt 170.7480  yes
## 207  59   male      typical 132.4184 300.4245 false      normal 170.7480  yes
## 208  60 female      typical 132.4184 300.4245 false      normal 170.7480   no
## 209  49   male  non-anginal 132.4184 300.4245 false         stt 170.7480   no
## 210  59   male      typical 132.4184 300.4245 false         stt 170.7480  yes
## 211  57   male  non-anginal 132.4184 300.4245 false      normal 170.7480   no
## 212  61   male      typical 132.4184 300.4245 false         stt 170.7480  yes
## 213  39   male      typical 132.4184 300.4245 false         stt 170.7480   no
## 214  61 female      typical 132.4184 300.4245 false      normal 170.7480  yes
## 215  56   male      typical 132.4184 300.4245  true      normal 170.7480  yes
## 216  43 female      typical 132.4184 300.4245  true      normal 170.7480  yes
## 217  62 female  non-anginal 132.4184 300.4245 false         stt 170.7480   no
## 218  63   male      typical 132.4184 300.4245  true      normal 170.7480  yes
## 219  65   male      typical 132.4184 300.4245 false      normal 170.7480   no
## 220  48   male      typical 132.4184 300.4245  true      normal 170.7480  yes
## 221  63 female      typical 132.4184 300.4245 false      normal 170.7480   no
## 222  55   male      typical 132.4184 300.4245 false         stt 170.7480  yes
## 223  65   male asymptomatic 132.4184 300.4245  true      normal 170.7480   no
## 224  56 female      typical 132.4184 300.4245  true      normal 170.7480  yes
## 225  54   male      typical 132.4184 300.4245 false         stt 170.7480  yes
## 226  70   male      typical 132.4184 300.4245 false         stt 170.7480  yes
## 227  62   male     atypical 132.4184 300.4245 false      normal 170.7480   no
## 228  35   male      typical 132.4184 300.4245 false         stt 170.7480  yes
## 229  59   male asymptomatic 132.4184 300.4245 false      normal 170.7480   no
## 230  64   male  non-anginal 132.4184 300.4245 false         stt 170.7480  yes
## 231  47   male  non-anginal 132.4184 300.4245 false         stt 170.7480   no
## 232  57   male      typical 132.4184 300.4245  true      normal 170.7480   no
## 233  55   male      typical 132.4184 300.4245 false      normal 170.7480  yes
## 234  64   male      typical 132.4184 300.4245 false      normal 170.7480  yes
## 235  70   male      typical 132.4184 300.4245 false      normal 170.7480   no
## 236  51   male      typical 132.4184 300.4245 false         stt 170.7480  yes
## 237  58   male      typical 132.4184 300.4245 false      normal 170.7480   no
## 238  60   male      typical 132.4184 300.4245 false      normal 170.7480   no
## 239  77   male      typical 157.0000 300.4245 false      normal 170.7480  yes
## 240  35   male      typical 157.0000 300.4245 false      normal 170.7480  yes
## 241  70   male  non-anginal 157.0000 300.4245 false         stt 170.7480  yes
## 242  59 female      typical 157.0000 300.4245 false         stt 170.7480  yes
## 243  64   male      typical 157.0000 300.4245 false      normal 170.7480   no
## 244  57   male      typical 157.0000 300.4245 false         stt 170.7480  yes
## 245  56   male      typical 157.0000 300.4245 false      normal 170.7480  yes
## 246  48   male      typical 157.0000 300.4245 false      normal 170.7480   no
## 247  56 female      typical 157.0000 300.4245 false      normal 170.7480  yes
## 248  66   male     atypical 157.0000 300.4245 false         stt 170.7480  yes
## 249  54   male     atypical 157.0000 300.4245 false      normal 170.7480   no
## 250  69   male  non-anginal 157.0000 300.4245 false      normal 170.7480   no
## 251  51   male      typical 157.0000 300.4245 false         stt 170.7480  yes
## 252  43   male      typical 157.0000 300.4245  true      normal 170.7480  yes
## 253  62 female      typical 157.0000 300.4245  true         stt 170.7480   no
## 254  67   male      typical 157.0000 300.4245 false      normal 170.7480  yes
## 255  59   male asymptomatic 157.0000 300.4245 false      normal 170.7480   no
## 256  45   male      typical 157.0000 300.4245 false      normal 170.7480  yes
## 257  58   male      typical 157.0000 300.4245 false      normal 170.7480  yes
## 258  50   male      typical 157.0000 300.4245 false      normal 170.7480  yes
## 259  62 female      typical 157.0000 300.4245 false         stt 170.7480  yes
## 260  38   male asymptomatic 157.0000 300.4245 false         stt 170.7480  yes
## 261  66 female      typical 157.0000 300.4245  true         stt 170.7480  yes
## 262  52   male      typical 157.0000 300.4245 false         stt 170.7480   no
## 263  53   male      typical 157.0000 300.4245 false         stt 170.7480  yes
## 264  63 female      typical 157.0000 300.4245 false         stt 170.7480  yes
## 265  54   male      typical 157.0000 300.4245 false      normal 170.7480  yes
## 266  66   male      typical 157.0000 300.4245 false      normal 170.7480  yes
## 267  55 female      typical 157.0000 300.4245 false hypertrophy 170.7480  yes
## 268  49   male  non-anginal 157.0000 300.4245 false      normal 170.7480   no
## 269  54   male      typical 157.0000 300.4245 false      normal 170.7480  yes
## 270  56   male      typical 157.0000 300.4245  true      normal 170.7480  yes
## 271  46   male      typical 157.0000 300.4245 false      normal 170.7480   no
## 272  61   male asymptomatic 157.0000 300.4245 false         stt 170.7480   no
## 273  67   male      typical 157.0000 300.4245 false         stt 170.7480   no
## 274  58   male      typical 157.0000 300.4245 false         stt 170.7480   no
## 275  47   male      typical 157.0000 300.4245 false      normal 170.7480  yes
## 276  52   male      typical 157.0000 300.4245 false         stt 170.7480   no
## 277  58   male      typical 157.0000 300.4245 false         stt 170.7480   no
## 278  57   male     atypical 157.0000 300.4245 false         stt 170.7480   no
## 279  58 female     atypical 157.0000 300.4245  true      normal 170.7480   no
## 280  61   male      typical 157.0000 300.4245 false      normal 170.7480  yes
## 281  42   male      typical 157.0000 300.4245 false         stt 170.7480  yes
## 282  52   male      typical 157.0000 300.4245  true         stt 170.7480  yes
## 283  59   male  non-anginal 157.0000 300.4245  true         stt 170.7480   no
## 284  40   male      typical 157.0000 300.4245 false         stt 170.7480   no
## 285  61   male      typical 157.0000 300.4245 false      normal 170.7480  yes
## 286  46   male      typical 157.0000 300.4245 false         stt 170.7480  yes
## 287  59   male asymptomatic 157.0000 300.4245 false         stt 170.7480   no
## 288  57   male     atypical 157.0000 300.4245 false      normal 170.7480   no
## 289  57   male      typical 157.0000 300.4245 false         stt 170.7480  yes
## 290  55 female      typical 157.0000 300.4245 false hypertrophy 170.7480  yes
## 291  61   male      typical 157.0000 300.4245 false         stt 170.7480   no
## 292  58   male      typical 157.0000 300.4245 false hypertrophy 170.7480   no
## 293  58 female      typical 157.0000 300.4245  true      normal 170.7480  yes
## 294  67   male  non-anginal 157.0000 300.4245 false      normal 170.7480   no
## 295  44   male      typical 157.0000 300.4245 false         stt 170.7480  yes
## 296  63   male      typical 157.0000 300.4245 false      normal 170.7480  yes
## 297  63 female      typical 157.0000 300.4245 false         stt 170.7480  yes
## 298  59   male      typical 157.0000 300.4245  true      normal 170.7480   no
## 299  57 female      typical 157.0000 300.4245 false         stt 170.7480  yes
## 300  45   male asymptomatic 157.0000 300.4245 false         stt 170.7480   no
## 301  68   male      typical 157.0000 300.4245  true         stt 170.7480   no
## 302  57   male      typical 157.0000 300.4245 false         stt 170.7480  yes
## 303  57 female     atypical 157.0000 300.4245 false      normal 170.7480   no
##     oldpeak  slp caa      thall output trtbps_bin chol_bin thalachh_bin
## 1       2.3 down   0     normal      1       high   medium       medium
## 2       3.5 down   0      fixed      1     medium   medium         high
## 3       1.4   up   0      fixed      1     medium      low         high
## 4       0.8   up   0      fixed      1        low   medium         high
## 5       0.6   up   0      fixed      1        low     high         high
## 6       0.4 flat   0     normal      1     medium      low       medium
## 7       1.3 flat   0      fixed      1     medium     high       medium
## 8       0.0   up   0 reversible      1        low     high         high
## 9       0.5   up   0 reversible      1       high      low         high
## 10      1.6   up   0      fixed      1       high      low         high
## 11      1.2   up   0      fixed      1     medium   medium         high
## 12      0.2   up   0      fixed      1     medium     high       medium
## 13      0.6   up   0      fixed      1     medium     high         high
## 14      1.8 flat   0      fixed      1        low      low       medium
## 15      1.0   up   0      fixed      1       high     high         high
## 16      1.6 flat   0      fixed      1        low      low       medium
## 17      0.0   up   0      fixed      1        low     high         high
## 18      2.6 down   0      fixed      1       high   medium          low
## 19      1.5   up   0      fixed      1       high   medium         high
## 20      1.8   up   2      fixed      1     medium   medium       medium
## 21      0.5 flat   0 reversible      1     medium   medium         high
## 22      0.4   up   0      fixed      1     medium   medium         high
## 23      0.0   up   0      fixed      1     medium   medium         high
## 24      1.0 flat   0      fixed      1       high   medium       medium
## 25      1.4   up   0 reversible      1     medium      low         high
## 26      0.4   up   2      fixed      1       high     high         high
## 27      1.6   up   0      fixed      1       high      low       medium
## 28      0.6   up   0      fixed      1        low      low       medium
## 29      0.8   up   1      fixed      1     medium     high       medium
## 30      1.2 down   0      fixed      1     medium      low       medium
## 31      0.0   up   1      fixed      1        low      low         high
## 32      0.4   up   0 reversible      1        low      low       medium
## 33      0.0   up   0      fixed      1     medium      low         high
## 34      0.5 down   1      fixed      1     medium     high       medium
## 35      1.4   up   1      fixed      1     medium      low       medium
## 36      1.4 down   0      fixed      1       high      low         high
## 37      0.0   up   0      fixed      1     medium     high         high
## 38      1.6   up   0 reversible      1       high   medium         high
## 39      0.8   up   0      fixed      1       high     high       medium
## 40      0.8   up   0      fixed      1       high     high       medium
## 41      1.5   up   1      fixed      1     medium     high       medium
## 42      0.2 flat   0      fixed      1     medium   medium         high
## 43      3.0 flat   0      fixed      1        low      low       medium
## 44      0.4 flat   0      fixed      1     medium     high       medium
## 45      0.0   up   0      fixed      1     medium     high         high
## 46      0.2   up   0      fixed      1        low     high         high
## 47      0.0   up   0      fixed      1     medium   medium         high
## 48      0.0   up   0      fixed      1     medium   medium       medium
## 49      0.0   up   0       none      1     medium      low       medium
## 50      0.0   up   0      fixed      1     medium   medium         high
## 51      0.5   up   0      fixed      1     medium   medium      medium
## 52      0.4 flat   0      fixed      1        low     high      medium
## 53      1.8 flat   3 reversible      1     medium   medium      medium
## 54      0.6 flat   0      fixed      1        low      low        high
## 55      0.0   up   0      fixed      1     medium   medium        high
## 56      0.8   up   1      fixed      1     medium      low      medium
## 57      0.0   up   0      fixed      1     medium   medium        high
## 58      0.0   up   0      fixed      1        low   medium        high
## 59      0.0   up   0      fixed      1        low      low        high
## 60      0.0   up   1      fixed      1     medium     high        high
## 61      0.0   up   1      fixed      1        low     high      medium
## 62      0.0   up   0 reversible      1        low     high      medium
## 63      0.0 flat   0     normal      1        low      low        high
## 64      0.0 flat   0     normal      1     medium      low      medium
## 65      0.0   up   0      fixed      1     medium      low        high
## 66      1.4   up   0      fixed      1     medium      low        high
## 67      1.2 flat   0      fixed      1        low   medium      medium
## 68      0.6 flat   0      fixed      1     medium   medium        high
## 69      0.0   up   0      fixed      1        low      low        high
## 70      0.0   up   0      fixed      1     medium      low        high
## 71      0.4 flat   0 reversible      1        low   medium      medium
## 72      0.0   up   1 reversible      1        low   medium      medium
## 73      0.0   up   0      fixed      1     medium      low        high
## 74      0.0   up   0      fixed      1     medium     high        high
## 75      0.2 flat   0      fixed      1     medium      low        high
## 76      1.4 flat   0      fixed      1     medium   medium        high
## 77      2.4 flat   0      fixed      1     medium   medium        high
## 78      0.0   up   0      fixed      1     medium   medium        high
## 79      0.0   up   0      fixed      1     medium      low        high
## 80      0.6 flat   0 reversible      1        low   medium      medium
## 81      0.0   up   0      fixed      1        low   medium        high
## 82      0.0   up   0      fixed      1     medium     high        high
## 83      0.0   up   1      fixed      1        low     high        high
## 84      1.2 flat   0 reversible      1       high     high        high
## 85      0.6 flat   0      fixed      1        low     high      medium
## 86      1.6 flat   0 reversible      1        low     high        high
## 87      1.0   up   1 reversible      1        low     high      medium
## 88      0.0   up   0 reversible      1        low      low      medium
## 89      1.6 flat   0      fixed      1        low      low      medium
## 90      1.0 flat   0      fixed      1        low   medium      medium
## 91      0.0   up   2      fixed      1     medium   medium        high
## 92      0.0   up   0 reversible      1     medium      low        high
## 93      0.0   up   4      fixed      1     medium   medium        high
## 94      0.0   up   1      fixed      1     medium     high        high
## 95      0.0 flat   0      fixed      1        low      low      medium
## 96      0.0   up   0 reversible      1       high   medium         low
## 97      1.2 flat   0      fixed      1     medium     high      medium
## 98      0.1   up   3 reversible      1        low   medium      medium
## 99      1.9   up   1      fixed      1     medium     high        high
## 100     0.0   up   3      fixed      1     medium   medium        high
## 101     0.8   up   2      fixed      1       high   medium        high
## 102     4.2 down   0 reversible      1       high     high      medium
## 103     0.0   up   2      fixed      1     medium      low        high
## 104     0.8 down   0 reversible      1        low   medium        high
## 105     0.0   up   0      fixed      1     medium      low         high
## 106     1.5 flat   0      fixed      1        low      low       medium
## 107     0.1 flat   1      fixed      1       high   medium       medium
## 108     0.2 flat   0      fixed      1     medium   medium       medium
## 109     1.1   up   0      fixed      1        low   medium         high
## 110     0.0   up   0      fixed      1        low   medium         high
## 111     0.0   up   0      fixed      1       high     high       medium
## 112     0.2   up   1 reversible      1       high      low         high
## 113     0.2   up   0 reversible      1     medium     high       medium
## 114     0.0   up   0 reversible      1        low      low         high
## 115     0.0   up   0      fixed      1     medium     high       medium
## 116     0.0   up   0      fixed      1        low      low         high
## 117     2.0 flat   0      fixed      1     medium      low         high
## 118     1.9 flat   0 reversible      1        low      low         high
## 119     0.0   up   0      fixed      1        low      low         high
## 120     0.0 flat   0      fixed      1     medium   medium       medium
## 121     2.0 flat   2      fixed      1     medium     high       medium
## 122     0.0   up   0      fixed      1     medium     high         high
## 123     0.0   up   0      fixed      1        low     high         high
## 124     0.0   up   0      fixed      1        low     high         high
## 125     0.0   up   0      fixed      1        low      low         high
## 126     0.7   up   0      fixed      1        low      low         high
## 127     0.1   up   0      fixed      1        low      low       medium
## 128     0.0   up   1      fixed      1       high     high         high
## 129     0.1 flat   0      fixed      1     medium      low         high
## 130     0.2   up   1      fixed      1        low     high       medium
## 131     0.0   up   1      fixed      1       high      low         high
## 132     0.0 flat   0      fixed      1     medium     high         high
## 133     0.0   up   0      fixed      1        low     high         high
## 134     0.0   up   0      fixed      1        low   medium       medium
## 135     0.0   up   0      fixed      1     medium     high         high
## 136     0.0   up   0      fixed      1     medium     high         high
## 137     0.0   up   0      fixed      1        low      low          low
## 138     0.0   up   0      fixed      1     medium      low       medium
## 139     1.5 flat   0     normal      1        low      low       medium
## 140     0.2 flat   1 reversible      1     medium     high          low
## 141     0.6   up   0      fixed      1        low     high       medium
## 142     1.2 flat   0      fixed      1        low     high         high
## 143     0.0 flat   0      fixed      1        low      low         high
## 144     0.3   up   2      fixed      1        low   medium       medium
## 145     1.1 flat   0      fixed      1     medium      low       medium
## 146     0.0   up   0      fixed      1       high   medium       medium
## 147     0.3 flat   1      fixed      1        low   medium       medium
## 148     0.9   up   0      fixed      1       high   medium         high
## 149     0.0   up   0      fixed      1        low   medium         high
## 150     0.0   up   0      fixed      1     medium      low       medium
## 151     2.3   up   0     normal      1       high   medium       medium
## 152     1.6 flat   0      fixed      1        low      low       medium
## 153     0.6 flat   0 reversible      1       high   medium       medium
## 154     0.0 flat   1      fixed      1       high     high       medium
## 155     0.0 flat   0      fixed      1     medium      low       medium
## 156     0.6 flat   0      fixed      1     medium      low       medium
## 157     0.0   up   0      fixed      1     medium   medium         high
## 158     0.0   up   0      fixed      1     medium      low         high
## 159     0.4 flat   4 reversible      1     medium      low       medium
## 160     0.0   up   0 reversible      1     medium   medium         high
## 161     0.0 down   0      fixed      1        low   medium         high
## 162     1.2   up   0      fixed      1     medium     high         high
## 163     0.0   up   0      fixed      1        low      low         high
## 164     0.0   up   4      fixed      1     medium      low         high
## 165     0.0   up   4      fixed      1     medium      low         high
## 166     1.5 flat   3      fixed      0       high     high          low
## 167     2.6 flat   2 reversible      0        low   medium       medium
## 168     3.6 down   2      fixed      0     medium     high         high
## 169     1.4 flat   1 reversible      0     medium   medium       medium
## 170     3.1 down   0 reversible      0     medium      low       medium
## 171     0.6 flat   1     normal      0     medium   medium       medium
## 172     1.0 down   0 reversible      0        low   medium         high
## 173     1.8 flat   0      fixed      0        low     high         high
## 174     3.2   up   2 reversible      0     medium   medium         high
## 175     2.4 flat   2 reversible      0     medium      low       medium
## 176     2.0 flat   0 reversible      0        low      low          low
## 177     1.4   up   2 reversible      0        low   medium         high
## 178     0.0   up   0      fixed      0     medium     high       medium
## 179     2.5 flat   0 reversible      0        low      low       medium
## 180     0.6 flat   1     normal      0       high     high          low
## 181     1.2 flat   1 reversible      0     medium     high       medium
## 182     1.0 flat   3 reversible      0       high   medium          low
## 183     0.0   up   0      fixed      0     medium     high         high
## 184     2.5 flat   1 reversible      0        low   medium         high
## 185     2.6 flat   0 reversible      0       high   medium       medium
## 186     0.0   up   1      fixed      0        low     high       medium
## 187     1.4   up   1 reversible      0     medium   medium       medium
## 188     2.2 flat   1 reversible      0     medium     high          low
## 189     0.6 flat   1 reversible      0     medium   medium         high
## 190     0.0   up   0 reversible      0        low      low       medium
## 191     1.2 flat   0 reversible      0     medium     high       medium
## 192     2.2 flat   3 reversible      0     medium      low       medium
## 193     1.4 flat   1 reversible      0        low      low          low
## 194     2.8 flat   2 reversible      0       high     high       medium
## 195     3.0 flat   0      fixed      0     medium      low       medium
## 196     3.4 down   0 reversible      0       high     high       medium
## 197     3.6 flat   0      fixed      0       high   medium       medium
## 198     0.2 flat   2 reversible      0     medium   medium         high
## 199     1.8 flat   2 reversible      0        low     high          low
## 200     0.6   up   2     normal      0        low   medium       medium
## 201     0.0   up   1      fixed      0        low      low         high
## 202     2.8 flat   1 reversible      0     medium   medium       medium
## 203     0.8   up   0 reversible      0       high     high          low
## 204     1.6 flat   0 reversible      0       high     high       medium
## 205     6.2 down   3 reversible      0       high      low       medium
## 206     0.0   up   1 reversible      0     medium   medium         high
## 207     1.2 flat   1 reversible      0        low   medium       medium
## 208     2.6 flat   2 reversible      0       high   medium       medium
## 209     2.0 flat   3 reversible      0        low      low       medium
## 210     0.0   up   1 reversible      0     medium      low         high
## 211     0.4 flat   1 reversible      0     medium   medium       medium
## 212     3.6 flat   1 reversible      0        low   medium       medium
## 213     1.2 flat   0 reversible      0        low    low       medium
## 214     1.0 flat   0 reversible      0       high   high       medium
## 215     1.2 flat   1      fixed      0     medium medium       medium
## 216     3.0 flat   0 reversible      0     medium   high       medium
## 217     1.2 flat   1 reversible      0     medium   high          low
## 218     1.8   up   3 reversible      0     medium   high       medium
## 219     2.8 flat   1 reversible      0     medium medium       medium
## 220     0.0   up   2 reversible      0     medium medium       medium
## 221     4.0 flat   3 reversible      0       high   high       medium
## 222     5.6 down   0 reversible      0     medium    low          low
## 223     1.4 flat   1      fixed      0     medium   high         high
## 224     4.0 down   2 reversible      0       high   high       medium
## 225     2.8 flat   1 reversible      0        low medium       medium
## 226     2.6 down   0 reversible      0       high    low       medium
## 227     1.4 flat   1 reversible      0        low   high          low
## 228     1.6 flat   0 reversible      0        low    low       medium
## 229     0.2 flat   0 reversible      0       high   high         high
## 230     1.8 flat   0 reversible      0     medium   high       medium
## 231     0.0   up   0      fixed      0        low medium       medium
## 232     1.0 flat   3 reversible      0       high   high       medium
## 233     0.8 flat   1 reversible      0       high   high       medium
## 234     2.2 down   1      fixed      0        low medium          low
## 235     2.4 flat   3      fixed      0     medium   high          low
## 236     1.6   up   0 reversible      0     medium   high         high
## 237     0.0   up   2 reversible      0     medium   high         high
## 238     1.2 flat   2 reversible      0     medium   high         high
## 239     0.0   up   3      fixed      0     medium   high         high
## 240     0.0   up   0 reversible      0     medium   high       medium
## 241     2.9 flat   1 reversible      0       high   high          low
## 242     0.0 flat   0      fixed      0       high medium       medium
## 243     2.0 flat   2     normal      0       high    low       medium
## 244     1.2 flat   1 reversible      0       high   high          low
## 245     2.1 flat   1     normal      0     medium    low          low
## 246     0.5 flat   0 reversible      0     medium   high         high
## 247     1.9 flat   2 reversible      0     medium   high       medium
## 248     0.0 flat   3     normal      0       high medium       medium
## 249     0.0   up   1 reversible      0       high   high         high
## 250     2.0 flat   3 reversible      0     medium medium       medium
## 251     4.2 flat   3 reversible      0     medium   high       medium
## 252     0.1 flat   4 reversible      0     medium medium       medium
## 253     1.9 flat   3      fixed      0     medium   high          low
## 254     0.9 flat   2      fixed      0        low   high       medium
## 255     0.0   up   0      fixed      0       high   high       medium
## 256     0.0 flat   3 reversible      0       high   high       medium
## 257     3.0 flat   2 reversible      0     medium medium       medium
## 258     0.9 flat   0 reversible      0       high    low       medium
## 259     1.4 flat   0      fixed      0       high medium       medium
## 260     3.8 flat   0 reversible      0        low medium         high
## 261     1.0 flat   2 reversible      0       high medium         high
## 262     0.0   up   1      fixed      0        low medium         high
## 263     2.0 flat   2 reversible      0     medium   high          low
## 264     1.8 flat   2      fixed      0        low   high         high
## 265     0.0 flat   1      fixed      0        low    low          low
## 266     0.1   up   1      fixed      0        low    low       medium
## 267     3.4 flat   0      fixed      0       high     high       medium
## 268     0.8   up   3      fixed      0        low      low       medium
## 269     3.2 flat   2      fixed      0     medium     high       medium
## 270     1.6 down   0 reversible      0     medium     high          low
## 271     0.8   up   0 reversible      0        low   medium       medium
## 272     2.6 flat   2      fixed      0     medium   medium       medium
## 273     1.0 flat   0      fixed      0        low   medium          low
## 274     0.1   up   1 reversible      0        low   medium       medium
## 275     1.0 flat   1      fixed      0        low     high       medium
## 276     1.0   up   2 reversible      0     medium      low         high
## 277     2.0 flat   1 reversible      0       high      low          low
## 278     0.3   up   0 reversible      0     medium     high       medium
## 279     0.0   up   2      fixed      0     medium     high       medium
## 280     3.6 flat   1      fixed      0     medium      low       medium
## 281     1.8 flat   0     normal      0     medium     high       medium
## 282     1.0 flat   0       none      0     medium      low       medium
## 283     2.2 flat   1     normal      0     medium      low       medium
## 284     0.0   up   0 reversible      0       high   medium         high
## 285     1.9   up   1 reversible      0     medium      low       medium
## 286     1.8 flat   2 reversible      0     medium     high       medium
## 287     0.8   up   2      fixed      0     medium      low         high
## 288     0.0   up   1      fixed      0       high   medium         high
## 289     3.0 flat   1 reversible      0        low     high       medium
## 290     2.0 flat   1 reversible      0     medium      low       medium
## 291     0.0   up   1 reversible      0       high      low         high
## 292     4.4 down   3     normal      0        low     high       medium
## 293     2.8 flat   2     normal      0       high   medium       medium
## 294     0.8 flat   0 reversible      0       high      low       medium
## 295     2.8 down   0     normal      0        low      low       medium
## 296     4.0   up   2 reversible      0     medium      low       medium
## 297     0.0 flat   0      fixed      0     medium      low       medium
## 298     1.0 flat   2     normal      0       high      low          low
## 299     0.2 flat   0 reversible      0     medium   medium       medium
## 300     1.2 flat   0 reversible      0        low     high       medium
## 301     3.4 flat   2 reversible      0       high      low       medium
## 302     1.2 flat   1 reversible      0     medium      low       medium
## 303     0.0 flat   1      fixed      0     medium   medium         high
```

Part e

```
################ Part e -- Clustering ################

# e1) subset df_heart by removing target and numerical features

clean_heart <- subset(df_heart, select = -c(output, trtbps, chol, thalachh))
head(clean_heart)
```

```
##    age    sex           cp   fbs restecg exng oldpeak  slp caa  thall trtbps_bin
## 1  63   male asymptomatic  true  normal   no     2.3 down   0 normal       high
## 2  37   male  non-anginal false     stt   no     3.5 down   0  fixed     medium
## 3  41 female     atypical false  normal   no     1.4   up   0  fixed     medium
## 4  56   male     atypical false     stt   no     0.8   up   0  fixed        low
## 5  57 female      typical false     stt  yes     0.6   up   0  fixed        low
## 6  57    male      typical false     stt    no     0.4 flat   0 normal     medium
##   chol_bin thalachh_bin
## 1   medium       medium
## 2   medium         high
## 3      low         high
## 4   medium         high
## 5     high         high
## 6      low       medium
```

```
# e2) Hierarchical clustering with Gower distance

# calculate distance
dist <- daisy(clean_heart, metric = "gower")
# hierarchical clustering
hc <- hclust(dist, method = "complete")
# plot dendrogram
plot(hc, labels = FALSE)
rect.hclust(hc, k = 4, border="red")
```



```
# choose k, number of clusters
cluster <- cutree(hc, k = 4)

# e3) MDS clustering with Manhattan distance

dist2 <- dist(clean_heart, method = "manhattan")
```

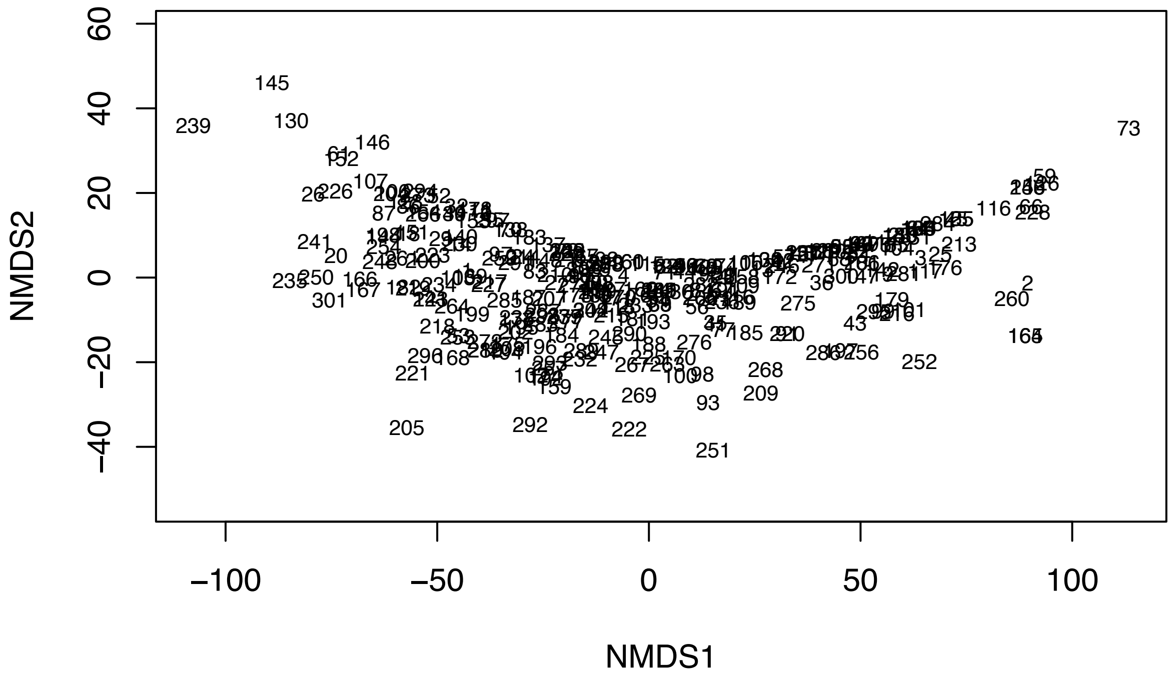
```
## Warning in dist(clean_heart, method = "manhattan"): NAs introduced by coercion
```

```
fit <- vegan::metaMDS(comm = dist2)
```

```
## Run 0 stress 0.05414163
## Run 1 stress 0.06033998
## Run 2 stress 0.05809318
## Run 3 stress 0.06441875
## Run 4 stress 0.05759696
## Run 5 stress 0.06076953
## Run 6 stress 0.06222502
## Run 7 stress 0.06699705
## Run 8 stress 0.07084782
## Run 9 stress 0.06375112
## Run 10 stress 0.05450698
## ... Procrustes: rmse 0.002298708  max resid 0.02711839
## Run 11 stress 0.06428604
## Run 12 stress 0.06565965
## Run 13 stress 0.0697716
## Run 14 stress 0.06468984
## Run 15 stress 0.07019834
## Run 16 stress 0.06331766
## Run 17 stress 0.06871108
## Run 18 stress 0.06789902
## Run 19 stress 0.07139322
## Run 20 stress 0.07047654
## *** No convergence -- monoMDS stopping criteria:
##      5: no. of iterations >= maxit
##      9: stress ratio > sratmax
##      6: scale factor of the gradient < sfgrmin
```

```
ordiplot(fit, type = "text")
```

```
## species scores not available
```



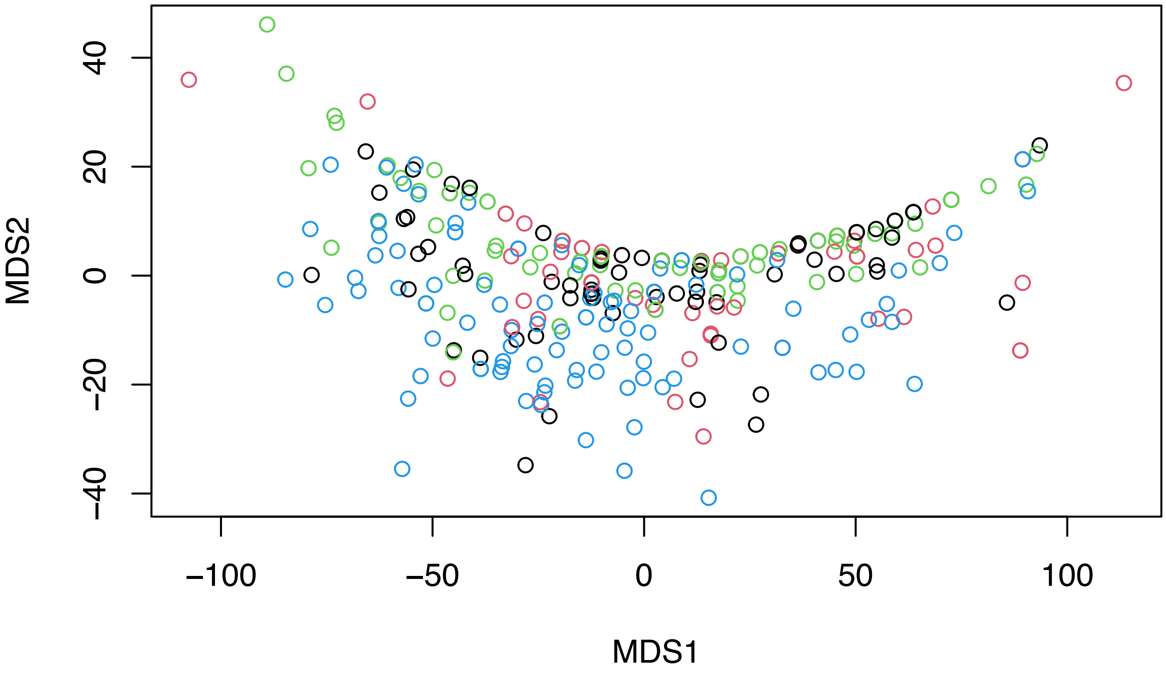
```
fit$stress
```

```
## [1] 0.05414163
```

```
# add cluster to original data
clean_heart <- cbind(clean_heart, cluster)
clean_heart$cluster <- factor(clean_heart$cluster)
head(clean_heart)
```

```
##   age    sex           cp   fbs restecg exng oldpeak  slp caa  thall trtbps_bin
## 1  63   male asymptomatic  true  normal   no     2.3 down   0 normal       high
## 2  37   male  non-anginal false     stt   no     3.5 down   0  fixed     medium
## 3  41 female     atypical false  normal   no     1.4   up   0  fixed     medium
## 4  56   male     atypical false     stt   no     0.8   up   0  fixed        low
## 5  57 female      typical false     stt  yes     0.6   up   0  fixed        low
## 6  57   male      typical false     stt   no     0.4 flat   0 normal     medium
##   chol_bin thalachh_bin cluster
## 1   medium       medium       1
## 2   medium         high       2
## 3      low         high       2
## 4   medium         high       1
## 5     high         high       3
## 6      low       medium       1
```

```r
# plot the clustering result with MDS
plot(fit$points, col = (clean_heart$cluster))
```



Part f

```r
################ Part f -- Classification ################

# f1) Decision Tree

# set seed and train control
set.seed(456)
train_control = trainControl(method = "cv", number = 10)

# prepare dataset for classification
lable_heart <- subset(df_heart, select = -c(trtbps, chol, thalachh))
head(lable_heart)
```

```
##   age    sex           cp   fbs restecg exng oldpeak  slp caa  thall output
## 1  63   male asymptomatic  true  normal   no     2.3 down   0 normal      1
## 2  37   male  non-anginal false     stt   no     3.5 down   0  fixed      1
## 3  41 female     atypical false  normal   no     1.4   up   0  fixed      1
## 4  56   male     atypical false     stt   no     0.8   up   0  fixed      1
## 5  57 female      typical false     stt  yes     0.6   up   0  fixed      1
## 6  57   male      typical false     stt   no     0.4 flat   0 normal      1
##   trtbps_bin chol_bin thalachh_bin
## 1       high   medium       medium
## 2     medium   medium         high
## 3     medium      low         high
## 4        low   medium         high
## 5        low     high         high
## 6     medium      low       medium
```

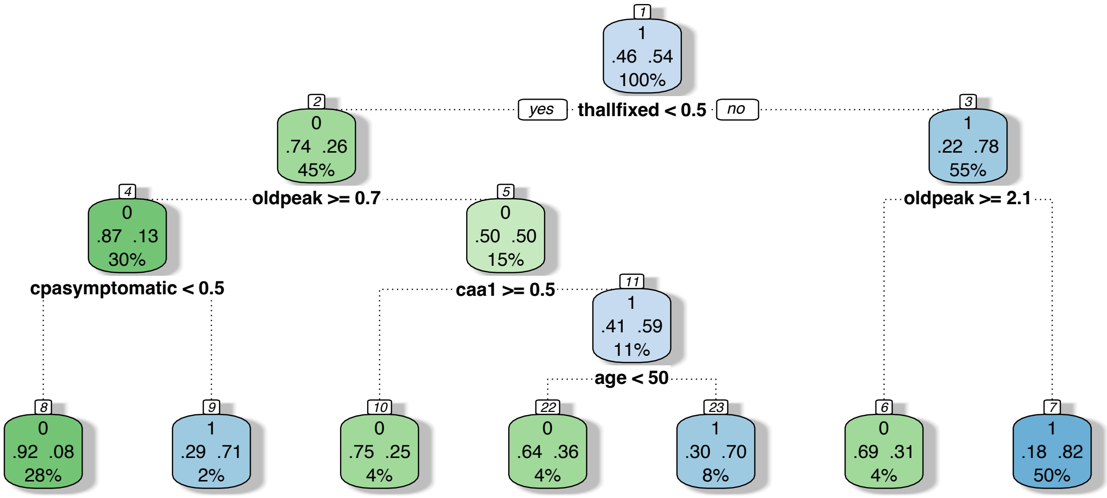
```
# fit the model
tree0 <- train(output ~., data = lable_heart, method = "rpart1SE", trControl = train_control)
tree0
```

```
## CART
##
## 303 samples
##  13 predictor
##   2 classes: ’0’, ’1’
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 272, 274, 273, 274, 272, 273, ...
## Resampling results:
##
##   Accuracy   Kappa
##   0.7188914  0.4306684
```

```
# evaluate the fit with a confusion matrix
pred_tree <- predict(tree0, lable_heart)

# confusion matrix
confusionMatrix(lable_heart$output, pred_tree)
```

```
## Confusion Matrix and Statistics
##
##             Reference
## Prediction   0   1
##          0 102  36
##          1  18 147
##
##                Accuracy : 0.8218
##                  95% CI : (0.774, 0.8632)
##     No Information Rate : 0.604
##     P-Value [Acc > NIR] : 2.35e-16
##
##                   Kappa : 0.6368
##
##  Mcnemar’s Test P-Value : 0.0207
##
##             Sensitivity : 0.8500
##             Specificity : 0.8033
##          Pos Pred Value : 0.7391
##          Neg Pred Value : 0.8909
##              Prevalence : 0.3960
##          Detection Rate : 0.3366
##    Detection Prevalence : 0.4554
##       Balanced Accuracy : 0.8266
##
##        'Positive' Class : 0
##
```

```
# visualize your decision tree0
fancyRpartPlot(tree0$finalModel, caption = "")
```



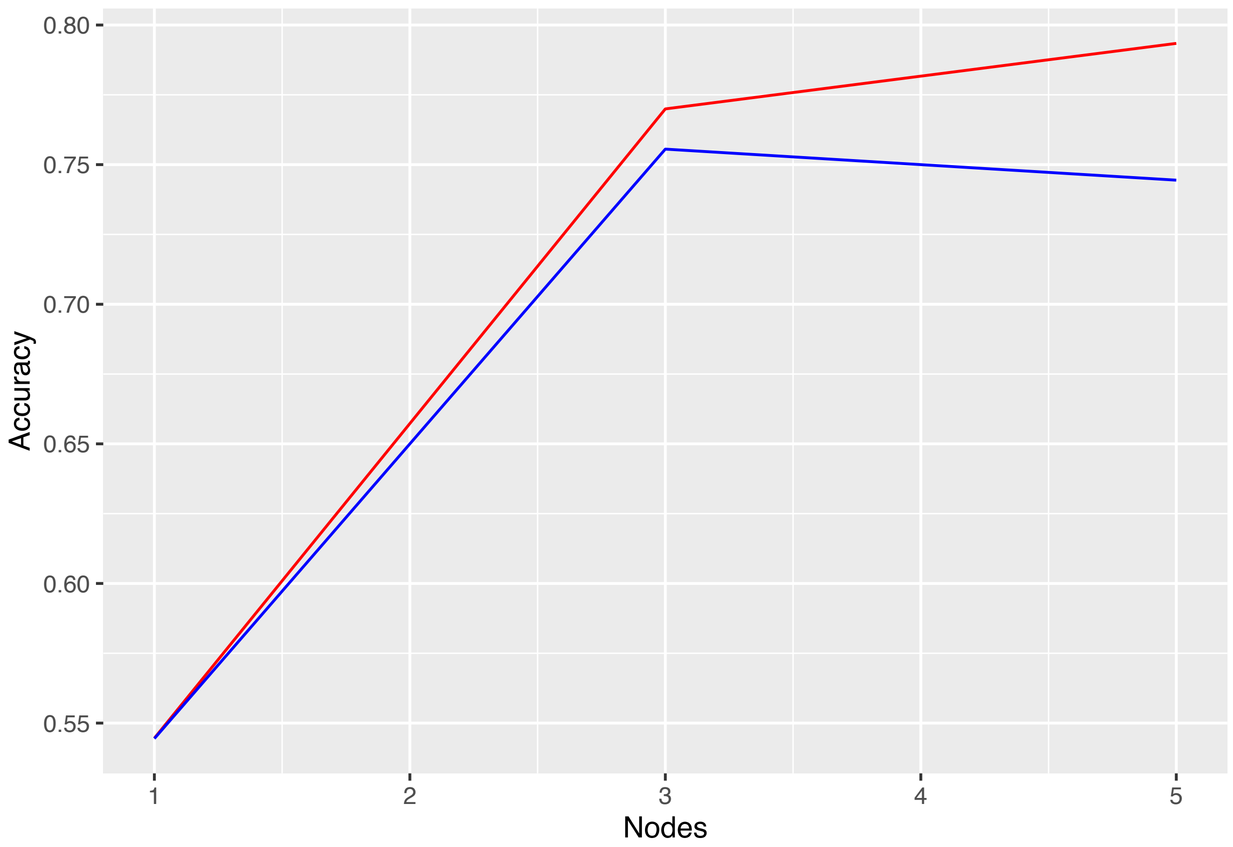
```
# validation (5 trees):

# partition the data
index = createDataPartition(y = lable_heart$output, p = 0.7, list = FALSE)
# everything in the generated index list
train_set = lable_heart[index,]
# everything except the generated indices
test_set = lable_heart[-index,]

# tree 1
hypers = rpart.control(minsplit = 2, maxdepth = 1, minbucket = 2)
tree1 <- train(output ~., data = train_set, control = hypers, trControl = train_control, method = "rpa
# train set
pred_tree <- predict(tree1, train_set)
cfm_train <- confusionMatrix(train_set$output, pred_tree)
# test set
```

```
pred_tree <- predict(tree1, test_set)
cfm_test <- confusionMatrix(test_set$output, pred_tree)
# training accuracy
a_train <- cfm_train$overall[1]
# testing accuracy
a_test <- cfm_test$overall[1]
# number of nodes
nodes <- nrow(tree1$finalModel$frame)

# form the table
comp_tbl <- data.frame("Nodes" = nodes, "TrainAccuracy" = a_train, "TestAccuracy" = a_test,
                       "MaxDepth" = 1, "Minsplit" = 2, "Minbucket" = 2)

# tree 2
hypers = rpart.control(minsplit = 5, maxdepth = 2, minbucket = 5)
tree2 <- train(output ~., data = train_set, control = hypers, trControl = train_control, method = "rpa
# training set
pred_tree <- predict(tree2, train_set)
cfm_train <- confusionMatrix(train_set$output, pred_tree)
# test set
pred_tree <- predict(tree2, test_set)
cfm_test <- confusionMatrix(test_set$output, pred_tree)
# training accuracy
a_train <- cfm_train$overall[1]
# testing accuracy
a_test <- cfm_test$overall[1]
# number of nodes
nodes <- nrow(tree2$finalModel$frame)

# add rows to the table - Make sure the order is correct
comp_tbl <- comp_tbl %>% rbind(list(nodes, a_train, a_test, 2, 5, 5))

# tree 3
hypers = rpart.control(minsplit = 20, maxdepth = 2, minbucket = 20)
tree3 <- train(output ~., data = train_set, control = hypers, trControl = train_control, method = "rpa
# training set
pred_tree <- predict(tree3, train_set)
cfm_train <- confusionMatrix(train_set$output, pred_tree)
# test set
pred_tree <- predict(tree3, test_set)
cfm_test <- confusionMatrix(test_set$output, pred_tree)
# training accuracy
a_train <- cfm_train$overall[1]
# testing accuracy
a_test <- cfm_test$overall[1]
# number of nodes
nodes <- nrow(tree3$finalModel$frame)

# add rows to the table - Make sure the order is correct
comp_tbl <- comp_tbl %>% rbind(list(nodes, a_train, a_test, 2, 20, 20))

# tree 4
hypers = rpart.control(minsplit = 40, maxdepth = 4, minbucket = 40)
```

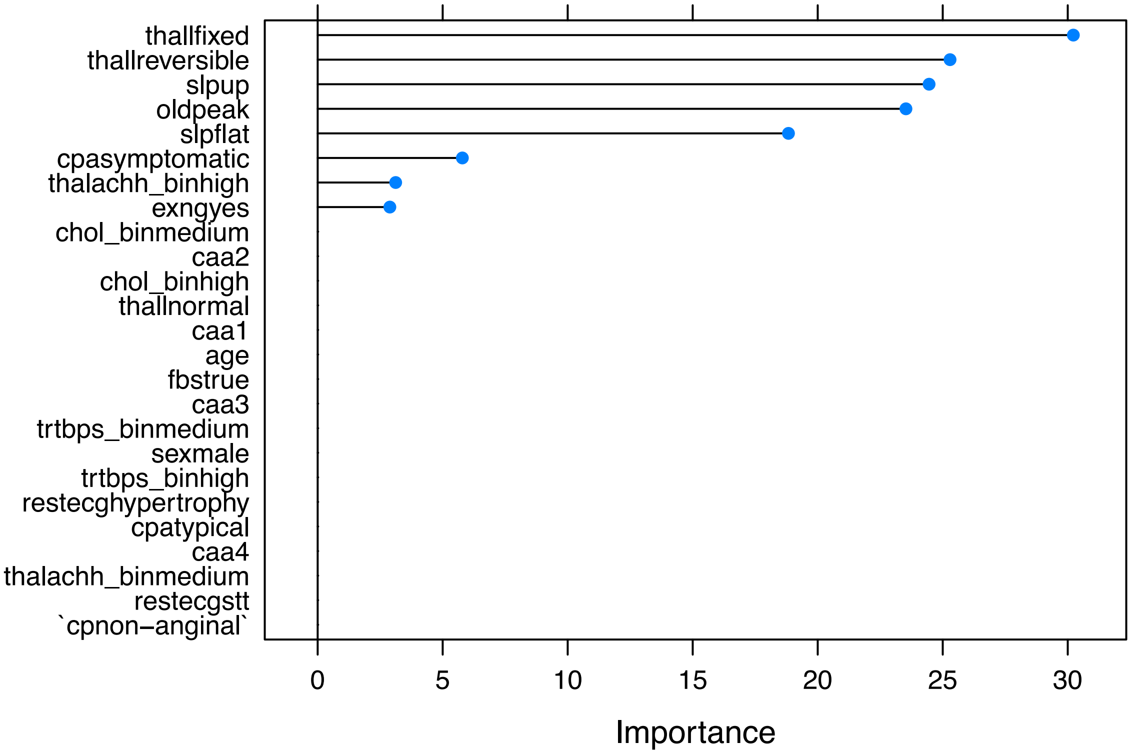
```r
tree4 <- train(output ~., data = train_set, control = hypers, trControl = train_control, method = "rpa
# training set
pred_tree <- predict(tree4, train_set)
cfm_train <- confusionMatrix(train_set$output, pred_tree)
# test set
pred_tree <- predict(tree4, test_set)
cfm_test <- confusionMatrix(test_set$output, pred_tree)
# training accuracy
a_train <- cfm_train$overall[1]
# testing accuracy
a_test <- cfm_test$overall[1]
# number of nodes
nodes <- nrow(tree3$finalModel$frame)

# add rows to the table - Make sure the order is correct
comp_tbl <- comp_tbl %>% rbind(list(nodes, a_train, a_test, 4, 40, 40))

# tree 5
hypers = rpart.control(minsplit = 500, maxdepth = 8, minbucket = 500)
tree5 <- train(output ~., data = train_set, control = hypers, trControl = train_control, method = "rpa
# training set
pred_tree <- predict(tree5, train_set)
cfm_train <- confusionMatrix(train_set$output, pred_tree)
# test set
pred_tree <- predict(tree5, test_set)
cfm_test <- confusionMatrix(test_set$output, pred_tree)
# training accuracy
a_train <- cfm_train$overall[1]
# testing accuracy
a_test <- cfm_test$overall[1]
# number of nodes
nodes <- nrow(tree5$finalModel$frame)

# add rows to the table - Make sure the order is correct
comp_tbl <- comp_tbl %>% rbind(list(nodes, a_train, a_test, 8, 500, 500))

# present table for comparison
comp_tbl
```

```
##          Nodes TrainAccuracy TestAccuracy MaxDepth Minsplit Minbucket
## Accuracy     3     0.7699531    0.7555556        1        2         2
## 1            5     0.7934272    0.7444444        2        5         5
## 11           3     0.7699531    0.7555556        2       20        20
## 12           3     0.7699531    0.7555556        4       40        40
## 13           1     0.5446009    0.5444444        8      500       500
```

```r
# visualize with line plot
ggplot(comp_tbl, aes(x=Nodes)) +
  geom_line(aes(y = TrainAccuracy), color = "red") +
  geom_line(aes(y = TestAccuracy), color="blue") +
  ylab("Accuracy")
```



```
# visualize the importance scores
importance <- varImp(tree2, scale = FALSE)
plot(importance)
```



```r
# f2) KNN

# work with a new data frame knn_heart
knn_heart <- subset(df_heart, select = -c(trtbps_bin, chol_bin, thalachh_bin))
head(knn_heart)
```

```
##   age    sex           cp   trtbps     chol   fbs restecg thalachh exng oldpeak
## 1  63   male asymptomatic 113.4639 194.6061  true  normal 103.8889   no     2.3
## 2  37   male  non-anginal 113.4639 194.6061 false     stt 103.8889   no     3.5
## 3  41 female     atypical 113.4639 194.6061 false  normal 103.8889   no     1.4
## 4  56   male     atypical 113.4639 194.6061 false     stt 103.8889   no     0.8
## 5  57 female      typical 113.4639 194.6061 false     stt 103.8889  yes     0.6
## 6  57   male      typical 113.4639 194.6061 false     stt 103.8889   no     0.4
##    slp caa  thall output
## 1 down   0 normal      1
## 2 down   0  fixed      1
## 3   up   0  fixed      1
## 4   up   0  fixed      1
## 5   up   0  fixed      1
## 6 flat   0 normal      1
```

```r
# convert categorical variables to dummies
dummy <- dummyVars(output ~ ., data = knn_heart)
# using the dummy predictor we need to transform our set into the dummy variable version
# the result won't be a data frame, so we need to transform it into one
dummies <- as.data.frame(predict(dummy, newdata = knn_heart))
```

```
## Warning in model.frame.default(Terms, newdata, na.action = na.action, xlev =
## object$lvls): variable 'output' is not a factor
```

```
# plug output back in to dummies
dummies$output <- df_heart$output
head(dummies)
```

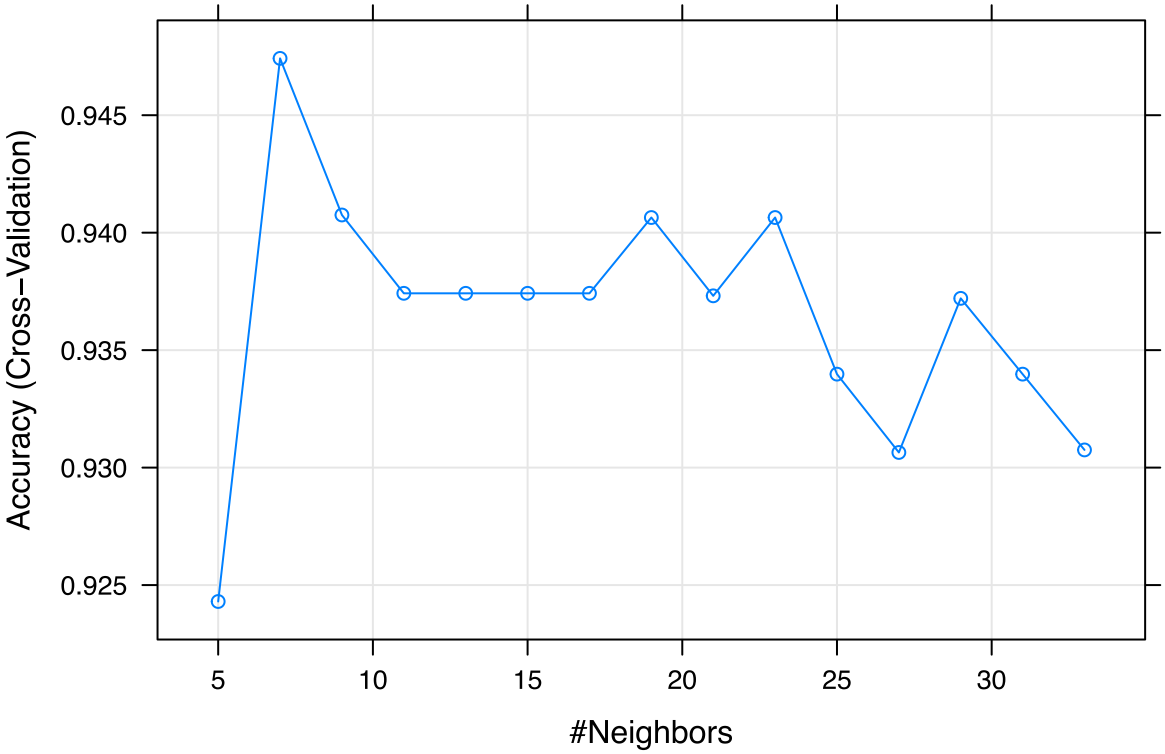
```
##   age sex.female sex.male cp.typical cp.atypical cp.non-anginal cp.asymptomatic
## 1  63          0        1          0           0              0               1
## 2  37          0        1          0           0              1               0
## 3  41          1        0          0           1              0               0
## 4  56          0        1          0           1              0               0
## 5  57          1        0          1           0              0               0
## 6  57          0        1          1           0              0               0
##     trtbps     chol fbs.false fbs.true restecg.normal restecg.stt
## 1 113.4639 194.6061         0        1              1           0
## 2 113.4639 194.6061         1        0              0           1
## 3 113.4639 194.6061         1        0              1           0
## 4 113.4639 194.6061         1        0              0           1
## 5 113.4639 194.6061         1        0              0           1
## 6 113.4639 194.6061         1        0              0           1
##   restecg.hypertrophy thalachh exng.no exng.yes oldpeak slp.down slp.flat
## 1                   0 103.8889       1        0     2.3        1        0
## 2                   0 103.8889       1        0     3.5        1        0
## 3                   0 103.8889       1        0     1.4        0        0
## 4                   0 103.8889       1        0     0.8        0        0
## 5                   0 103.8889       0        1     0.6        0        0
## 6                   0 103.8889       1        0     0.4        0        1
##   slp.up caa.0 caa.1 caa.2 caa.3 caa.4 thall.none thall.normal thall.fixed
## 1      0     1     0     0     0     0          0            1           0
## 2      0     1     0     0     0     0          0            0           1
## 3      1     1     0     0     0     0          0            0           1
## 4      1     1     0     0     0     0          0            0           1
## 5      1     1     0     0     0     0          0            0           1
## 6      0     1     0     0     0     0          0            1           0
##   thall.reversible output
## 1                0      1
## 2                0      1
## 3                0      1
## 4                0      1
## 5                0      1
## 6                0      1
```

```
# run the general knn
set.seed(123)
ctrl <- trainControl(method="cv", number = 10)
knnFit <- train(output ~ ., data = dummies,
                method = "knn",
                trControl = ctrl,
                preProcess = c("center","scale"),
                tuneLength = 15)
```

```
knnFit
```

```
## k-Nearest Neighbors
##
## 303 samples
##  30 predictor
##   2 classes: '0', '1'
##
## Pre-processing: centered (30), scaled (30)
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 272, 273, 272, 273, 273, 273, ...
## Resampling results across tuning parameters:
##
##   k   Accuracy   Kappa
##    5  0.9243011  0.8470702
##    7  0.9474194  0.8931076
##    9  0.9407527  0.8794698
##   11  0.9374194  0.8728070
##   13  0.9374194  0.8729888
##   15  0.9374194  0.8727280
##   17  0.9374194  0.8727280
##   19  0.9406452  0.8796264
##   21  0.9373118  0.8730187
##   23  0.9406452  0.8795631
##   25  0.9339785  0.8661008
##   27  0.9306452  0.8594931
##   29  0.9372043  0.8728612
##   31  0.9339785  0.8662237
##   33  0.9307527  0.8594139
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was k = 7.
```

```
plot(knnFit)
```



```r
# setup a tuneGrid with the tuning parameters
tuneGrid <- expand.grid(kmax = 3:7,                        # test a range of k values 3 to 7
                        kernel = c("rectangular", "cos"),  # regular and cosine-based distance functio
                        distance = 1:3)                    # powers of Minkowski 1 to 3

# tune and fit the model with 10-fold cross validation,
# standardization, and our specialized tune grid
kknn_fit <- train(output ~ .,
                  data = dummies,
                  method = 'kknn',
                  trControl = ctrl,
                  preProcess = c('center', 'scale'),
                  tuneGrid = tuneGrid)

# printing trained model provides report
kknn_fit
```

```
## k-Nearest Neighbors
##
## 303 samples
##  30 predictor
##   2 classes: '0', '1'
##
## Pre-processing: centered (30), scaled (30)
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 273, 272, 273, 273, 272, 273, ...
## Resampling results across tuning parameters:
##
##   kmax  kernel       distance  Accuracy   Kappa
##   3     rectangular  1         0.9374194  0.8732353
##   3     rectangular  2         0.8948387  0.7882205
##   3     rectangular  3         0.9047312  0.8077597
##   3     cos          1         0.9343011  0.8665642
##   3     cos          2         0.8880645  0.7726044
##   3     cos          3         0.8816129  0.7604443
##   4     rectangular  1         0.9374194  0.8732353
##   4     rectangular  2         0.9015054  0.8017340
##   4     rectangular  3         0.9047312  0.8077597
##   4     cos          1         0.9341935  0.8662436
##   4     cos          2         0.8946237  0.7867136
##   4     cos          3         0.8848387  0.7672380
##   5     rectangular  1         0.9408602  0.8797988
##   5     rectangular  2         0.9080645  0.8143204
##   5     rectangular  3         0.9113978  0.8211641
##   5     cos          1         0.9374194  0.8730534
##   5     cos          2         0.8947312  0.7875510
##   5     cos          3         0.8815054  0.7603912
##   6     rectangular  1         0.9408602  0.8797988
##   6     rectangular  2         0.9177419  0.8332540
##   6     rectangular  3         0.9113978  0.8211641
##   6     cos          1         0.9406452  0.8794436
##   6     cos          2         0.8880645  0.7741582
##   6     cos          3         0.8781720  0.7538436
##   7     rectangular  1         0.9440860  0.8866888
##   7     rectangular  2         0.9210753  0.8400397
##   7     rectangular  3         0.9179570  0.8337565
##   7     cos          1         0.9406452  0.8794436
##   7     cos          2         0.8947312  0.7876722
##   7     cos          3         0.8915054  0.7813043
##
## Accuracy was used to select the optimal model using the largest value.
## The final values used for the model were kmax = 7, distance = 1 and kernel
##  = rectangular.
```

```r
# knn prediction
pred_knn <- predict(kknn_fit, dummies)
# generate confusion matrix
confusionMatrix(dummies$output, pred_knn)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 130   8
##          1   3 162
##
##                Accuracy : 0.9637
##                  95% CI : (0.936, 0.9817)
##     No Information Rate : 0.5611
##     P-Value [Acc > NIR] : <2e-16
```

```
##
##                   Kappa : 0.9266
##
##  Mcnemar's Test P-Value : 0.2278
##
##             Sensitivity : 0.9774
##             Specificity : 0.9529
##          Pos Pred Value : 0.9420
##          Neg Pred Value : 0.9818
##              Prevalence : 0.4389
##          Detection Rate : 0.4290
##    Detection Prevalence : 0.4554
##       Balanced Accuracy : 0.9652
##
##        'Positive' Class : 0
##
```

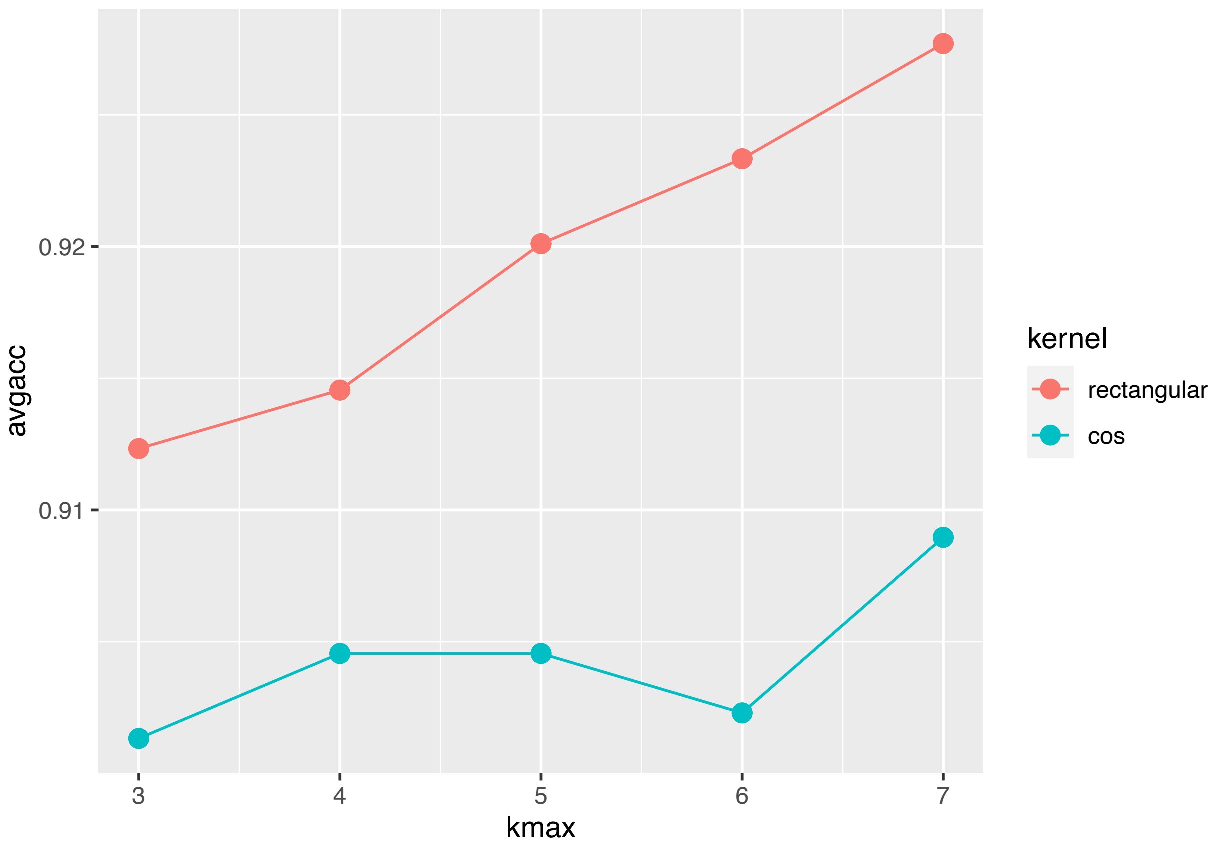
```
# gives just the table of results by parameter
knn_results <- kknn_fit$results
head(knn_results)
```

```
##   kmax      kernel distance  Accuracy     Kappa AccuracySD    KappaSD
## 1    3 rectangular        1 0.9374194 0.8732353 0.02815298 0.05856935
## 4    3         cos        1 0.9343011 0.8665642 0.05083405 0.10415912
## 2    3 rectangular        2 0.8948387 0.7882205 0.06036804 0.12200058
## 5    3         cos        2 0.8880645 0.7726044 0.06583567 0.13461393
## 3    3 rectangular        3 0.9047312 0.8077597 0.06717768 0.13570763
## 6    3         cos        3 0.8816129 0.7604443 0.06343266 0.12849914
```

```
# group by k and distance function, create an aggregation by averaging
knn_results <- knn_results %>%
  group_by(kmax, kernel) %>%
  mutate(avgacc = mean(Accuracy))
head(knn_results)
```

```
## # A tibble: 6 x 8
## # Groups:   kmax, kernel [2]
##    kmax kernel      distance Accuracy Kappa AccuracySD KappaSD avgacc
##   <int> <fct>          <int>    <dbl> <dbl>      <dbl>   <dbl>  <dbl>
## 1     3 rectangular        1    0.937 0.873     0.0282  0.0586  0.912
## 2     3 cos                1    0.934 0.867     0.0508  0.104   0.901
## 3     3 rectangular        2    0.895 0.788     0.0604  0.122   0.912
## 4     3 cos                2    0.888 0.773     0.0658  0.135   0.901
## 5     3 rectangular        3    0.905 0.808     0.0672  0.136   0.912
## 6     3 cos                3    0.882 0.760     0.0634  0.128   0.901
```

```
# plot aggregated (over Minkowski power) accuracy per k, split by distance function
ggplot(knn_results, aes(x=kmax, y=avgacc, color=kernel)) +
  geom_point(size=3) + geom_line()
```



Part g

```
################ Part g -- Evaluation ################

# select KNN as the better classifier

# g1) 2*2 confusion matrix

cm <- confusionMatrix(dummies$output, pred_knn)
cm
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 130   8
##          1   3 162
##
##                Accuracy : 0.9637
##                  95% CI : (0.936, 0.9817)
##     No Information Rate : 0.5611
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.9266
##
##  Mcnemar's Test P-Value : 0.2278
```

```
##
##             Sensitivity : 0.9774
##             Specificity : 0.9529
##          Pos Pred Value : 0.9420
##          Neg Pred Value : 0.9818
##              Prevalence : 0.4389
##          Detection Rate : 0.4290
##    Detection Prevalence : 0.4554
##       Balanced Accuracy : 0.9652
##
##        'Positive' Class : 0
##
```

```
# scoring metrics
metrics <- as.data.frame(cm$byClass)
# view the object
metrics
```

```
##                      cm$byClass
## Sensitivity           0.9774436
## Specificity           0.9529412
## Pos Pred Value        0.9420290
## Neg Pred Value        0.9818182
## Precision             0.9420290
## Recall                0.9774436
## F1                    0.9594096
## Prevalence            0.4389439
## Detection Rate        0.4290429
## Detection Prevalence  0.4554455
## Balanced Accuracy     0.9651924
```

```
# g2) calculate the precision and recall manually
# precision: TP/(TP+FP) = 130/(130+8) = 0.942
# recall: TP/(TP + FN) = 130/(130+3) = 0.977

# g3) produce ROC plot

# check target class and make sure it has 2 levels
str(dummies$output)
```

```
##  Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
```

```
# partition the data
index = createDataPartition(y=dummies$output, p=0.7, list=FALSE)
# everything in the generated index list
train_pima = dummies[index,]
# everything except the generated indices
test_pima = dummies[-index,]

# set control parameter
train_control = trainControl(method = "cv", number = 10)
# fit the model
```

```r
knn_pima <- train(output ~., data = train_pima, method = "knn", trControl = train_control, tuneLength
# evaluate fit
knn_pima
```

```
## k-Nearest Neighbors
##
## 213 samples
##  30 predictor
##   2 classes: ’0’, ’1’
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 192, 191, 192, 192, 192, 193, ...
## Resampling results across tuning parameters:
##
##   k   Accuracy   Kappa
##    5  0.9389394  0.8753587
##    7  0.9432468  0.8835357
##    9  0.9432468  0.8835357
##   11  0.9432468  0.8835357
##   13  0.9432468  0.8835357
##   15  0.9432468  0.8835357
##   17  0.9384848  0.8733949
##   19  0.9334848  0.8627634
##   21  0.9287229  0.8523287
##   23  0.9241775  0.8430850
##   25  0.9191775  0.8322297
##   27  0.9191775  0.8319123
##   29  0.9053247  0.8036792
##   31  0.9005628  0.7940022
##   33  0.9051082  0.8034026
##   35  0.9005628  0.7940022
##   37  0.8960173  0.7846018
##   39  0.8960173  0.7846018
##   41  0.8960173  0.7846018
##   43  0.8960173  0.7846018
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was k = 15.
```

```r
# evaluate the fit with a confusion matrix
pred_pima <- predict(knn_pima, test_pima)
# confusion Matrix
confusionMatrix(test_pima$output, pred_pima)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 38  3
##          1  0 49
##
```

```
##                Accuracy : 0.9667
##                  95% CI : (0.9057, 0.9931)
##     No Information Rate : 0.5778
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.9324
##
##  Mcnemar's Test P-Value : 0.2482
##
##             Sensitivity : 1.0000
##             Specificity : 0.9423
##          Pos Pred Value : 0.9268
##          Neg Pred Value : 1.0000
##              Prevalence : 0.4222
##          Detection Rate : 0.4222
##    Detection Prevalence : 0.4556
##       Balanced Accuracy : 0.9712
##
##        'Positive' Class : 0
##
```

```
# get class probabilities for KNN
pred_prob <- predict(knn_pima, test_pima, type = "prob")
head(pred_prob)
```

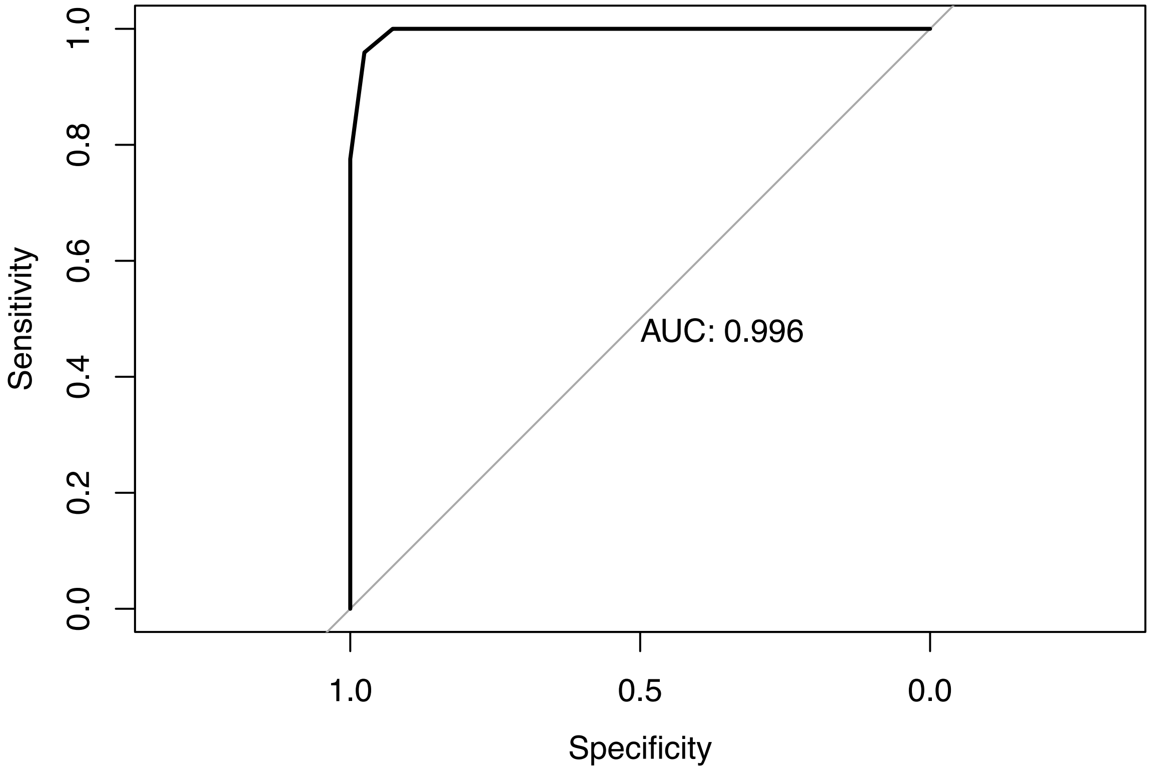
```
##   0 1
## 1 0 1
## 2 0 1
## 3 0 1
## 4 0 1
## 5 0 1
## 6 0 1
```

```
# create an ROC curve for our model.
roc_obj <- roc((test_pima$output), pred_prob[,1])
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls > cases
```

```
plot(roc_obj, print.auc=TRUE)
```

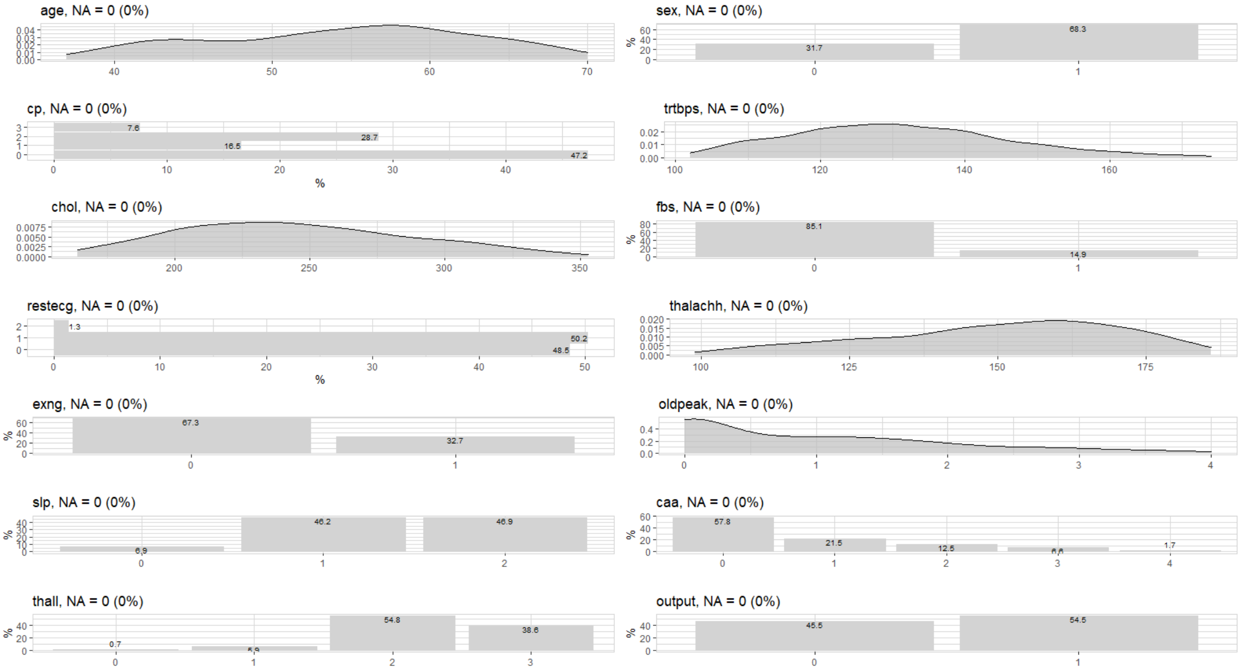


```
# The AUC value reaches almost 1.0 in this cases.
# The previous KNN model has 92.6% Kappa accuracy.
# The previous KNN model has 0.9637 general accuracy.
# Only 3 observations are miss-classified as FP.
# The ROC plot proves that this model is almost 100% accurate.
```

Heart Disease Report and Reflection

Addition

Before reporting the takeaways from heart disease data, it is important to add the missing part in section b2). Due to the size of the output, the overall exploration result was not generated completely in section b2) and the complete result is added here as a reference.



As an overview of this data, it is obvious to see that this dataset does not include any missing values and it contains mixed type of variables (categorical and numerical).

Report

As the dramatic growing of data science application in medical field, lots of valuable medical data are being studied in academia. This heart disease dataset is one of the disease data sets collected by UCI machine learning laboratory. By loading and exploring it, it presents relatively cleaned results such as 0 missing values, few influential and outlier points. It is always critical to acknowledge that having a relatively less messy data set saves a ton of time and improves analytics reliability in further study.

As a data set with mixed data types, converting them to the correct type plays an important role. Due to different analytical techniques, data type is one of the major step that would cause problem in analysis. PCA is one of the dimensionality reduction technique that only takes numerical values. However, due to the over all feature of heart data, PCA is not implemented since over 70% of the variables are categorical. However, it is still necessary to convert ordinal variables to factors and present necessary labels for those levels. For instance most of the categorical features are not only converted to factors but are also rewritten as factors with levels such as variable “chest pain” with leveled label typical angina, atypical angina, non-anginal pain and asymptomatic instead of only level numbers.

The pre-processing part took a lot of time in the whole process and this phase builds a valuable and concrete foundation for further machine learning output. The major work in this step is removing the noise of numerical variables and also smoothing the data. Normalization, binning, and smoothing are all applied in this phase. The completed binning and smoothing result on variable trtbps, chol and thalachh are all added to the complete df_heart dataset. The df_heart data set includes a total of 17 variables which are the 14 original variables with smoothed and binned values for trtbps, chol and thalachh but also trtbps_bin, chol_bin and thalachh_bin added as the addition columns. Building such clean and modified version based on the original data set successfully helps to manipulate the data set such as sub-setting new data frames or converting new data frames to dummies. This complete version is the main data frame in this project that can be re-used and applied to classifiers conveniently.

The clustering process causes some problems in the beginning and the reason is because the data type problem. Due to the overall categorical data type in this data set, K-means clustering would not work and generate results. Thus, hierarchical clustering and MDS are used in this step. Based on the hierarchical dendrogram and MDS clustering plot, it presents 2 main branches in the dendrogram which indicates the target variable (no heart disease as 0 or heart disease as 1). With MDS, it also generates out an up side down bell shaped curve. The possible reason that it does not have separated clusters might be the distance used and MDS itself.

By applying decision tree and KNN, we learn that KNN performs a lot better than decision tree. With 5 different trees in part g, it shows that best results of Kappa Accuracy only reaches 76% overall, however, the Kappa Accuracy in KNN reaches 92%. The precision and recall value are calculated over 92%. The further evaluation step proves that the accuracy in the experiment is reliable. The ROC curve of with KNN has AUC with 0.996. By relating to the confusion matrix on test set, it only shows 3 misclassified observation as false positive. It strongly demonstrates that our heart disease prediction model might be 100% correct.

One of the most interesting findings can be the performance of our KNN model. It is unbelievable to see that KNN beats decision tree with extraordinary performance. The reason of choosing decision tree as is because it presents tree plots and handles qualitative data as well. However, the result shows that KNN outperforms decision tree on this heart dataset. The reason might be because trees do not handle data as robustly as KNN can which causes problem in prediction. The other reason might be that KNN is better at handling rare cases especially in medical fields. In cancer research case studies, even though that many cancer cases are rare, but KNN generates better predictions than decision tree since decision trees usually prune important classes out of the model if there exists minority groups.

Reflection

Key terms such as data science, data mining and data analysis have grown as a prevalent topic that people always talk about nowadays. Before diving into the topic of data mining, I used to believe that data mining is a course about using techniques retrieve data just like “mining” bitcoin. However, I was absolutely wrong after I took this course. I realized that it is a huge topic that contains tons of topics from various fields. Data science was never an independent subject before, but it is now becoming an independent and interdisciplinary field.

By looking through, it is obvious to acknowledge how much data impacts the world. From simple linear regression model to complex classification algorithms, these are the fundamental tools for problem-solving. Advanced pre-processing skills such as normalization, binning and smoothing are found to be necessary steps for analysis. My knowledge stayed at the level of detecting outliers and influential points before taking this class but now I do have an understanding of using techniques to manipulate the given data. The major two areas in this course are supervised learning and unsupervised learning and I do believe that they are not only the important in this course but also plays an important role in data science foundation. Projecting data with PCA and then using K-means to cluster is such a creative way of visualizing and presenting data. Besides this, applying metrics to evaluate our classification results is also valuable. There is a ton of things that I have learned but the most important thing I have learned about this course is the impact of data science.

Three months ago, I read the paper published by Google's Deep Mind about its official 1.0 version of automatic code generator. I was extremely shocked by this technology because I couldn't believe how much such data science technology could change the world. My first impression of machine learning stayed at the time when AlphaGo dominated the "go" world and I had no idea how that even worked and now I am confident to hold a belief to demonstrate how much this is going to change the world. In last September, it was a great chance of joining Dr. Raicu's seminar to preview advanced research such as their cancer detection topic. By relating it to what I have done to this heart project, it is critical to understand the tremendous impact of data science to medical field. A model with 100% detection accuracy meaning no need for any additional human annotation and detection. In severe and common human diseases, such technology has an infinite potential. Obstacles such as limited information of data becomes a problem. Patients' lung nodule CT scans with not enough observation was also presented in Raicu's seminar. This is just one of the numerous challenges that people face which makes this filed with such a huge potential for us to explore.

By comparing the traditional software development field and data science field, it shows that even though traditional computer science field still has a high demanding in market, this demand has now become over-saturated and hiring freeze in Silicon Valley becomes normal. The big difference is that SDE has reached its limit of developing "vertically" and now it is developing "horizontally" where new subjects such as AI, robotics, and DS become independent fields. Programming is a strong ability that people should acquire in work, but it might soon be replaced advanced technology. It is said from Chinese ministry of education that python has become a required course in all primary schools. This means programming will only be considered as a must-have ability and independent fields like data science will soon take over many new fields in the future. I am glad that I have started my journey of learning this valuable subject and I also believe I should always be learning data tools to solve problems regardless their specific fields.