Project 1

2022-11-01

INCLUDING LIBRARY AND PLAYING AROUND WITH DATA

```
library(ggplot2)
library(forcats)
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --

## v tibble  3.1.8      v purrr   0.3.5
## v tidyr   1.2.1      v dplyr   1.0.10
## v readr   2.1.3      v stringr 1.4.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
##
## Attaching package: 'pROC'
##
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
##
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(grid)
library(ggplot2)
library(lattice)
```

```
data <- read.csv("heart.csv")
head(data)
```

```
##   age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal
## 1  63   1  3      145  233   1       0     150     0     2.3     0  0    1
## 2  37   1  2      130  250   0       1     187     0     3.5     0  0    2
## 3  41   0  1      130  204   0       0     172     0     1.4     2  0    2
## 4  56   1  1      120  236   0       1     178     0     0.8     2  0    2
## 5  57   0  0      120  354   0       1     163     1     0.6     2  0    2
## 6  57   1  0      140  192   0       1     148     0     0.4     1  0    1
##   target
## 1      1
## 2      1
## 3      1
## 4      1
## 5      1
## 6      1
```

```
summary(data)
```

```
##       age             sex               cp           trestbps
##  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          thalach
##  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
##      exang           oldpeak         slope             ca
##  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
##       thal           target
##  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
```

```
glimpse(data)
```

```
## Rows: 303
## Columns: 14
## $ age      <int> 63, 37, 41, 56, 57, 57, 56, 44, 52, 57, 54, 48, 49, 64, 58, 5~
## $ sex      <int> 1, 1, 0, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 1, 0, 0, 0, 0, 1, 0, 1~
## $ cp       <int> 3, 2, 1, 1, 0, 0, 1, 1, 2, 2, 0, 2, 1, 3, 3, 2, 2, 3, 0, 3, 0~
## $ trestbps <int> 145, 130, 130, 120, 120, 140, 140, 120, 172, 150, 140, 130, 1~
## $ chol     <int> 233, 250, 204, 236, 354, 192, 294, 263, 199, 168, 239, 275, 2~
## $ fbs      <int> 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0~
## $ restecg  <int> 0, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1~
## $ thalach  <int> 150, 187, 172, 178, 163, 148, 153, 173, 162, 174, 160, 139, 1~
## $ exang    <int> 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0~
## $ oldpeak  <dbl> 2.3, 3.5, 1.4, 0.8, 0.6, 0.4, 1.3, 0.0, 0.5, 1.6, 1.2, 0.2, 0~
## $ slope    <int> 0, 0, 2, 2, 2, 1, 1, 2, 2, 2, 2, 2, 2, 1, 2, 1, 2, 0, 2, 2, 1~
## $ ca       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2~
## $ thal     <int> 1, 2, 2, 2, 2, 1, 2, 3, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3~
## $ target   <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
```

```
ncol(data)
```

```
## [1] 14
```

```
nrow(data)
```

```
## [1] 303
```

```
colnames(data)
```

```
##  [1] "age"      "sex"      "cp"       "trestbps" "chol"     "fbs"
##  [7] "restecg"  "thalach"  "exang"    "oldpeak"  "slope"    "ca"
## [13] "thal"     "target"
```

```
summary(data)
```

```
##       age             sex               cp           trestbps
##  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          thalach
##  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
##      exang           oldpeak         slope             ca
```

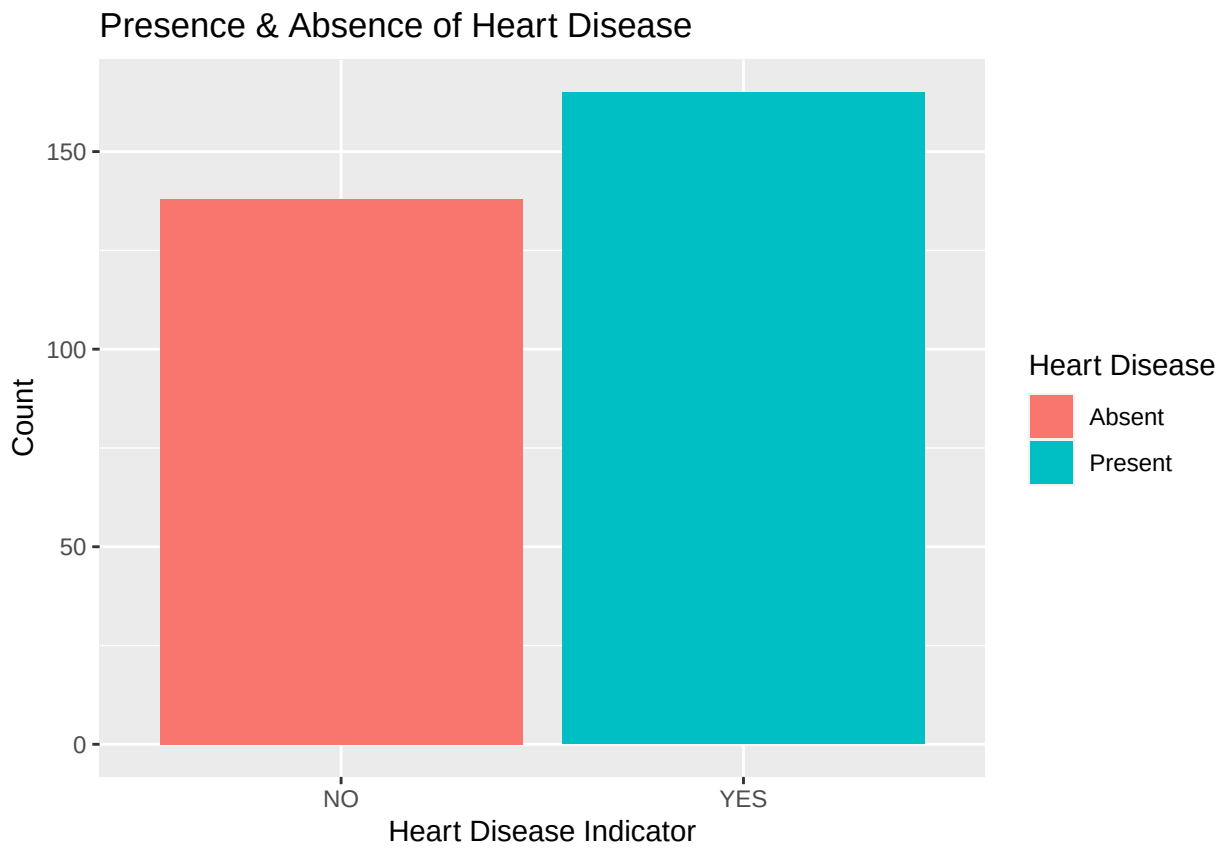
```
##  Min.   :0.0000   Min.   :0.00   Min.   :0.000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.00   1st Qu.:1.000   1st Qu.:0.0000
##  Median :0.0000   Median :0.80   Median :1.000   Median :0.0000
##  Mean   :0.3267   Mean   :1.04   Mean   :1.399   Mean   :0.7294
##  3rd Qu.:1.0000   3rd Qu.:1.60   3rd Qu.:2.000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :6.20   Max.   :2.000   Max.   :4.0000
##       thal           target
##  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
```

DATA TRANSFORMATION

```
data2 <- data %>%
     mutate(sex = if_else(sex == 1, "MALE", "FEMALE"),
            fbs = if_else(fbs == 1, ">120", "<=120"),
            exang = if_else(exang == 1, "YES" ,"NO"),
            cp = if_else(cp == 1, "ATYPICAL ANGINA",
                         if_else(cp == 2, "NON-ANGINAL PAIN", "ASYMPTOMATIC")),
            restecg = if_else(restecg == 0, "NORMAL",
                              if_else(restecg == 1, "ABNORMALITY", "PROBABLE OR DEFINITE")),
            slope = as.factor(slope),
            ca = as.factor(ca),
            thal = as.factor(thal),
            target = if_else(target == 1, "YES", "NO")
     ) %>%
     mutate_if(is.character, as.factor) %>%
     dplyr::select(target, sex, fbs, exang, cp, restecg, slope, ca, thal, everything())
```

DATA VISUALIZATION

```
ggplot(data2, aes(x=target, fill=target))+
     geom_bar()+
     xlab("Heart Disease Indicator")+
     ylab("Count")+
     ggtitle("Presence & Absence of Heart Disease")+
     scale_fill_discrete(name= 'Heart Disease', labels =c("Absent", "Present"))
```



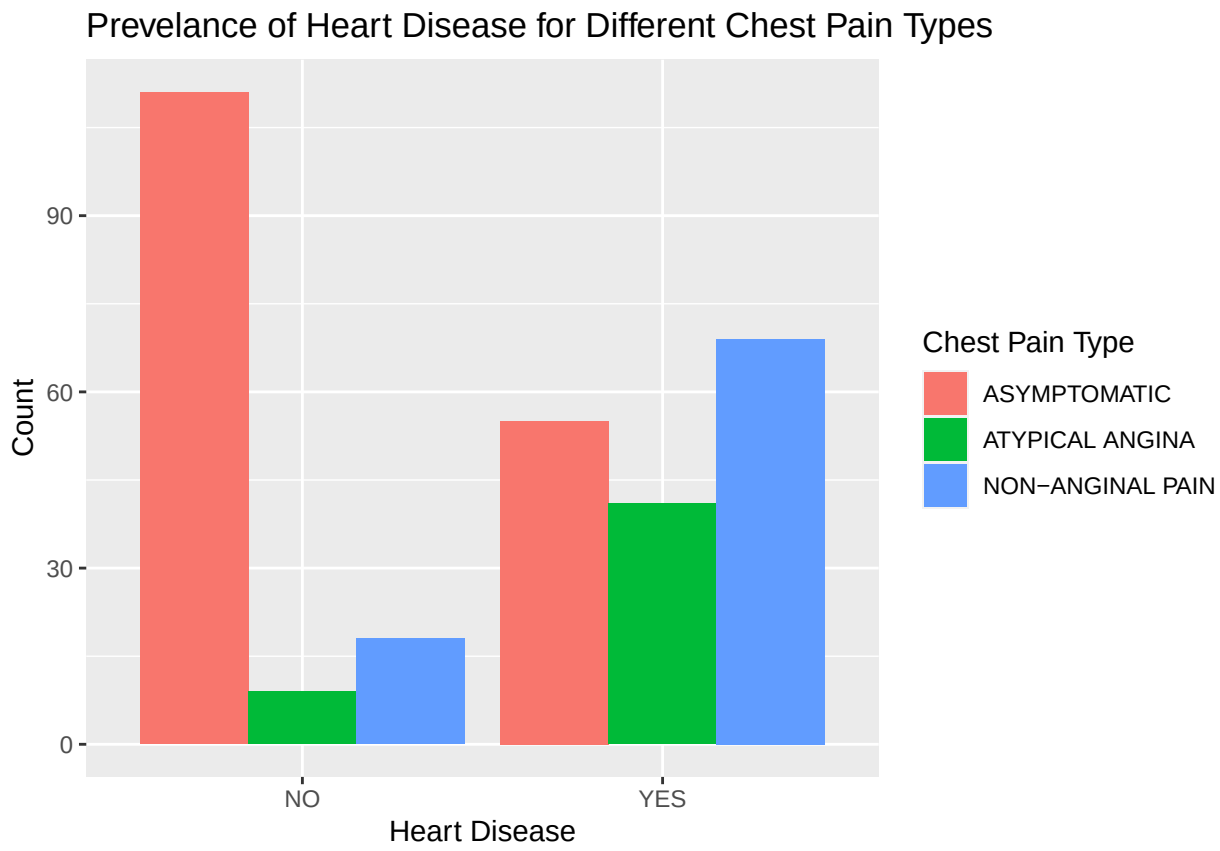

```
age.plot <- ggplot(data2, mapping = aes(x = age, fill = target)) +
  stat_count(binwidth=0.5) +
  facet_wrap(vars(target)) +
  labs(title = "Prevelance of Heart Disease Across Age", x = "Age (years)", y = "Count", fill = "Heart
```

```
## Warning: Ignoring unknown parameters: binwidth
```

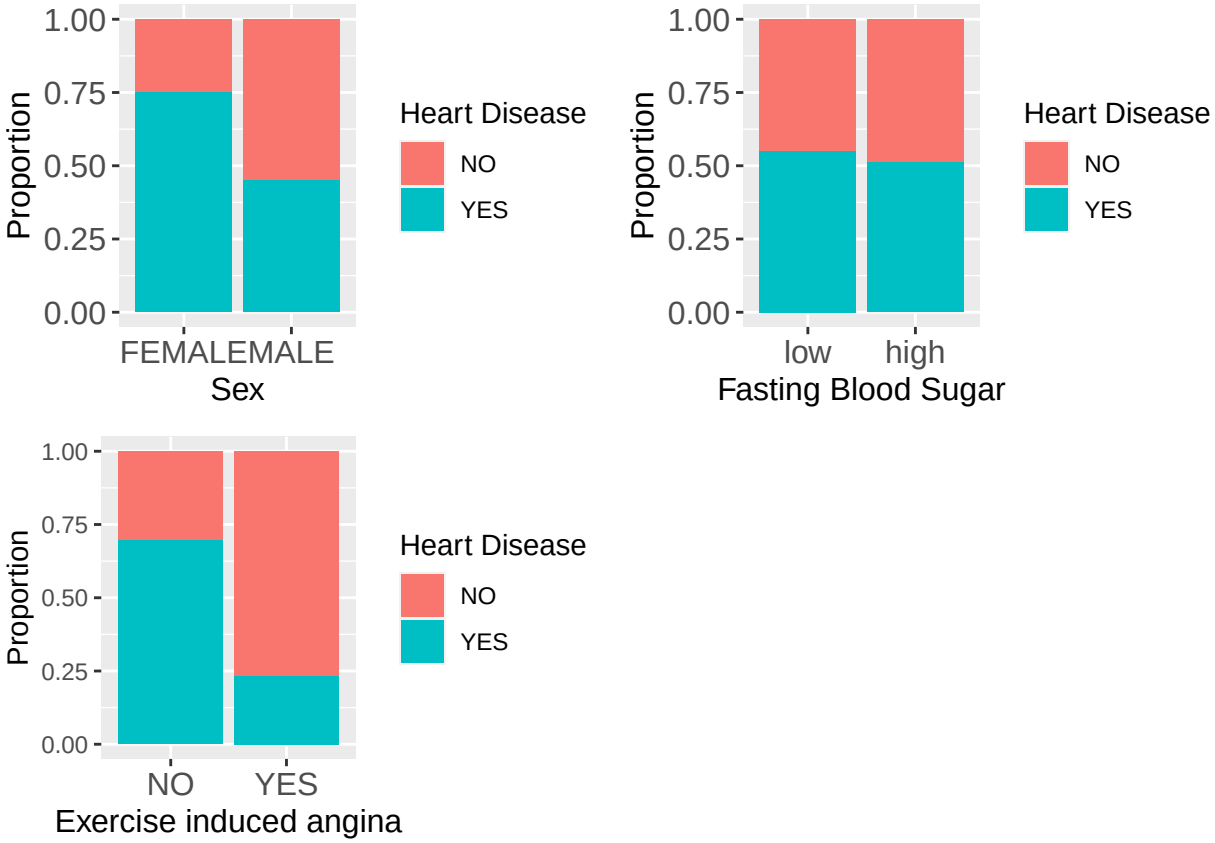
```
age.plot
```


```
cp.plot <- ggplot(data2, mapping = aes(x=target, fill = cp)) +
  geom_bar(position = "dodge") +
  labs(title = "Prevelance of Heart Disease for Different Chest Pain Types", x = "Heart Disease", y = "
```

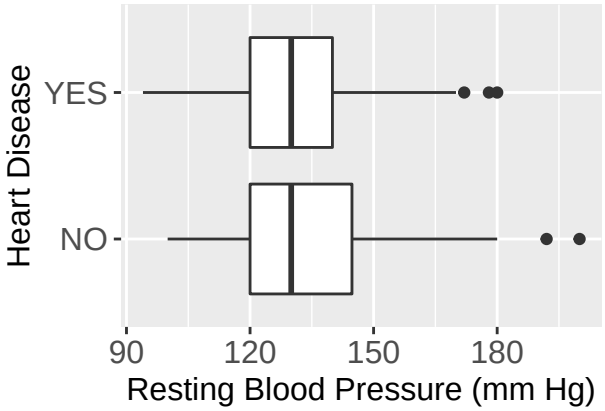
```
cp.plot
```

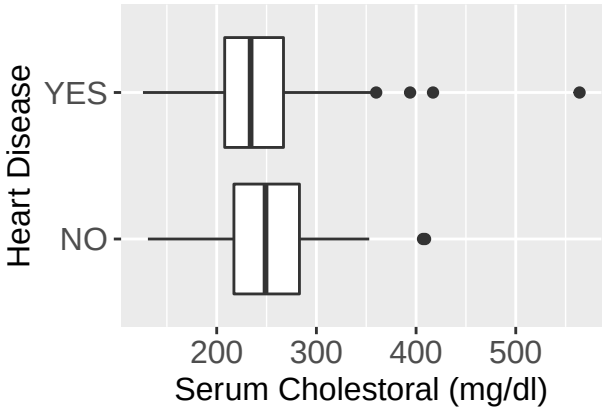


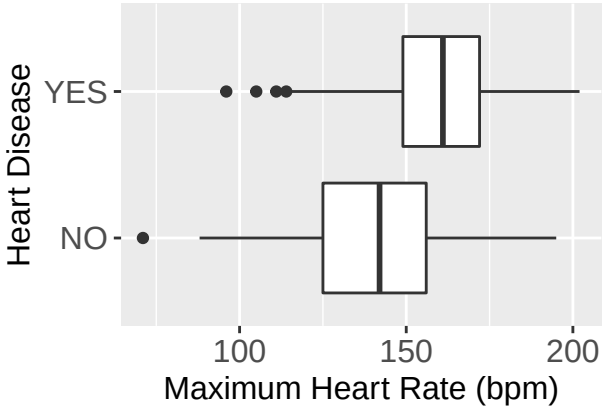

```
sex.plot <- ggplot(data2, mapping = aes(x = sex, fill = target)) +
  geom_bar(position = "fill") +
  labs(x = "Sex", y = "Proportion", fill = "Heart Disease") +
  theme(axis.text.x = element_text(size = 12), axis.title.x = element_text(size = 12),
        axis.title.y = element_text(size = 12), axis.text.y = element_text(size = 12))

fbs.plot <- ggplot(data2, mapping = aes(x=fbs, fill=target)) +
  geom_bar(position = "fill") +
  labs(x = "Fasting Blood Sugar", y = "Proportion", fill = "Heart Disease") +
  scale_x_discrete(labels = c("low", "high"))+
  theme(axis.text.x = element_text(size = 12), axis.title.x = element_text(size = 12),
        axis.title.y = element_text(size = 12), axis.text.y = element_text(size = 12))

exang.plot <- ggplot(data2, mapping = aes(x = exang, fill = target)) +
  geom_bar(position = "fill") +
  labs(x = "Exercise induced angina", y = "Proportion", fill = "Heart Disease") +
  theme(axis.text.x = element_text(size = 12), axis.title.x = element_text(size = 12))

grid.arrange(sex.plot, fbs.plot, exang.plot, nrow=2)
```




```
trestbps.plot <- ggplot(data2, mapping = aes(x=trestbps, y=target)) +
  geom_boxplot() +
  labs(x = "Resting Blood Pressure (mm Hg)", y = "Heart Disease") +
  theme(axis.text.x = element_text(size = 12), axis.title.x = element_text(size = 12),
        axis.title.y = element_text(size = 12), axis.text.y = element_text(size = 12))

chol.plot <- ggplot(data2, mapping = aes(x=chol, y=target)) +
  geom_boxplot() +
  labs(x = "Serum Cholestoral (mg/dl)", y = "Heart Disease") +
  theme(axis.text.x = element_text(size = 12), axis.title.x = element_text(size = 12),
        axis.title.y = element_text(size = 12), axis.text.y = element_text(size = 12))

thalach.plot <- ggplot(data2, mapping = aes(x = thalach, y = target)) +
  geom_boxplot() +
  labs(x = "Maximum Heart Rate (bpm)", y = "Heart Disease") +
  theme(axis.text.x = element_text(size = 12), axis.title.x = element_text(size = 12),
        axis.title.y = element_text(size = 12), axis.text.y = element_text(size = 12))

grid.arrange(trestbps.plot, chol.plot, thalach.plot, nrow=2)
```





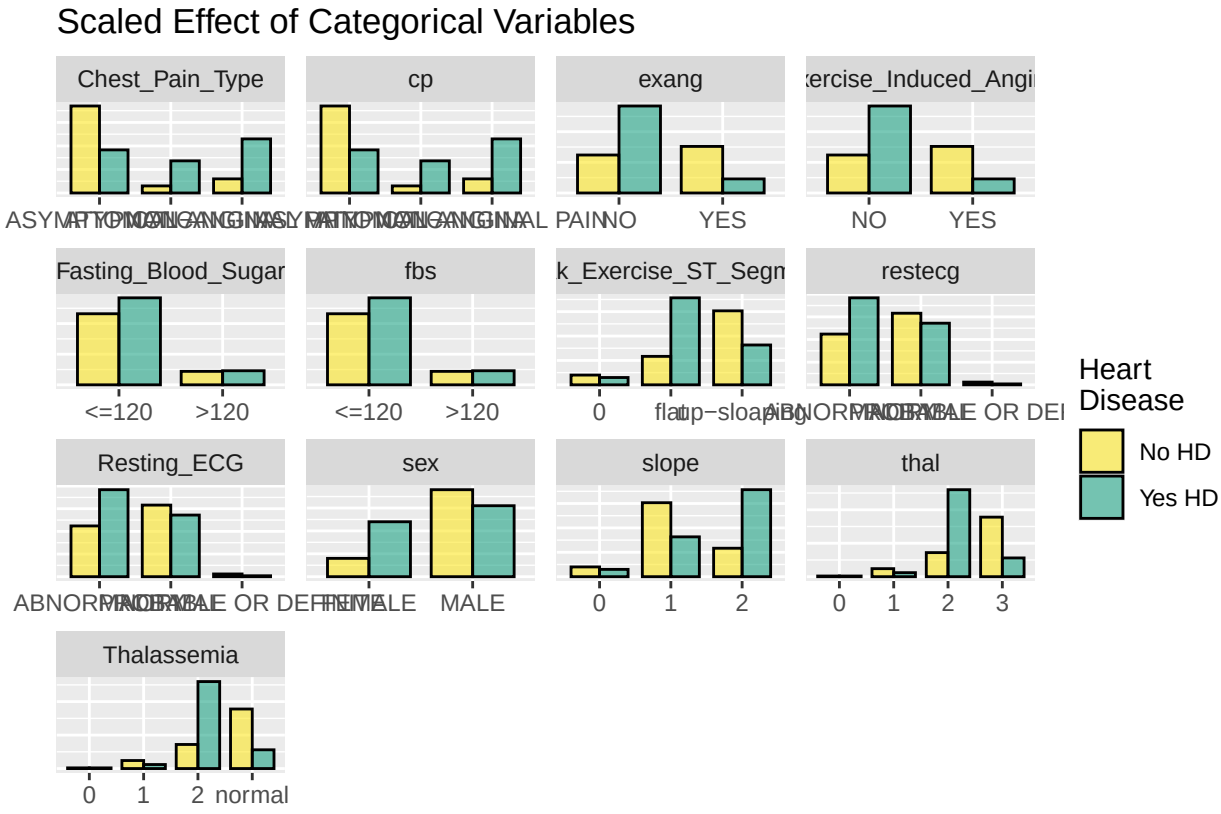



```
#Select categorical vars, recode them to their character values, convert to long format
data3 <- data2 %>%
  select(sex,
         cp,
         fbs,
         restecg,
         exang,
         slope,
         thal,
         target) %>%
  mutate(sex = recode_factor(sex, `0` = "female",
                                  `1` = "male" ),
         Chest_Pain_Type = recode_factor(cp, `1` = "typical",
                                             `2` = "atypical",
                                             `3` = "non-angina",
                                             `4` = "asymptomatic"),
         Fasting_Blood_Sugar = recode_factor(fbs, `0` = "<= 120 mg/dl",
                                                  `1` = "> 120 mg/dl"),
         Resting_ECG = recode_factor(restecg, `0` = "normal",
                                              `1` = "ST-T abnormality",
                                              `2` = "LV hypertrophy"),
         Exercise_Induced_Angina = recode_factor(exang, `0` = "no",
                                                        `1` = "yes"),
         Peak_Exercise_ST_Segment = recode_factor(slope, `1` = "up-sloaping",
                                                         `2` = "flat",
                                                         `3` = "down-sloaping"),
```

```
        Thalassemia = recode_factor(thal, `3` = "normal",
                                          `6` = "fixed defect",
                                          `7` = "reversible defect")) %>%
  gather(key = "key", value = "value", -target)
```

```
## Warning: attributes are not identical across measure variables;
## they will be dropped
```

```
#Visualize with bar plot
data3 %>%
  ggplot(aes(value)) +
    geom_bar(aes(x        = value,
                 fill     = target),
                 alpha    = .6,
                 position = "dodge",
                 color    = "black",
                 width    = .8
             ) +
    labs(x = "",
         y = "",
         title = "Scaled Effect of Categorical Variables") +
    theme(
         axis.text.y  = element_blank(),
         axis.ticks.y = element_blank()) +
    facet_wrap(~ key, scales = "free", nrow = 4) +
    scale_fill_manual(
         values = c("#fde725ff", "#20a486ff"),
         name   = "Heart\nDisease",
         labels = c("No HD", "Yes HD"))
```

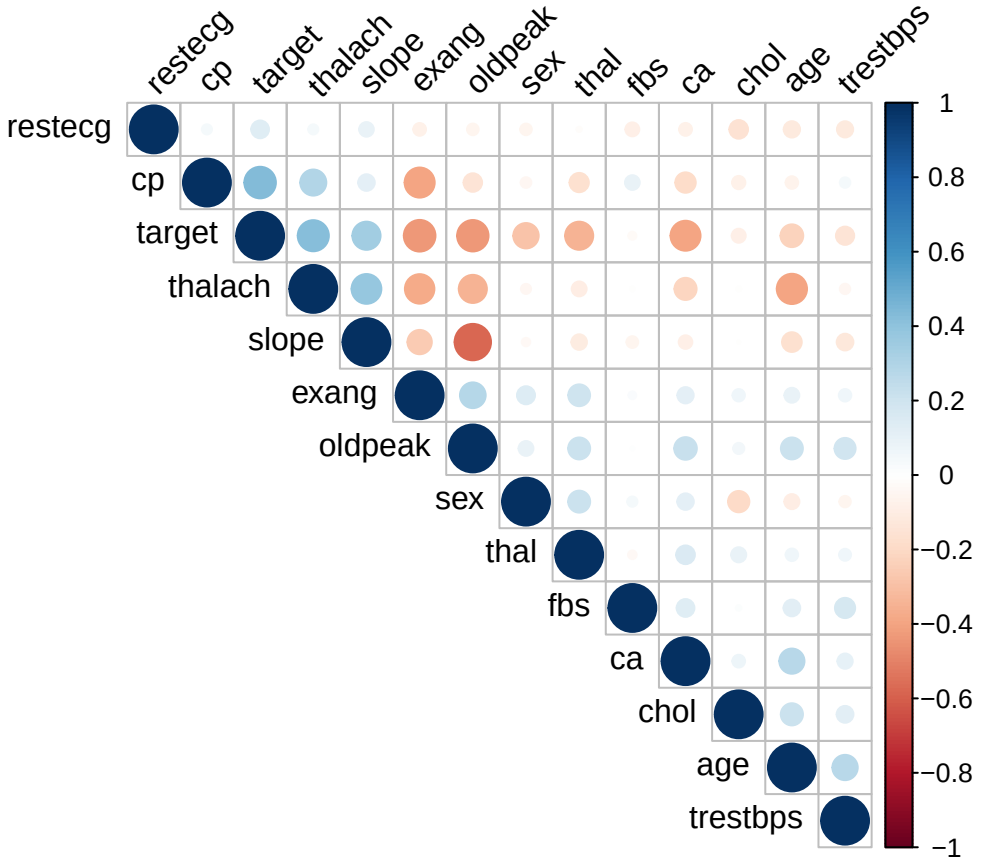



```
cor_heart <- cor(data[, 1:14])
cor_heart
```

```
##                 age         sex          cp    trestbps         chol
## age       1.00000000 -0.09844660 -0.06865302  0.27935091  0.213677957
## sex      -0.09844660  1.00000000 -0.04935288 -0.05676882 -0.197912174
## cp       -0.06865302 -0.04935288  1.00000000  0.04760776 -0.076904391
## trestbps  0.27935091 -0.05676882  0.04760776  1.00000000  0.123174207
## chol      0.21367796 -0.19791217 -0.07690439  0.12317421  1.000000000
## fbs       0.12130765  0.04503179  0.09444403  0.17753054  0.013293602
## restecg  -0.11621090 -0.05819627  0.04442059 -0.11410279 -0.151040078
## thalach  -0.39852194 -0.04401991  0.29576212 -0.04669773 -0.009939839
## exang     0.09680083  0.14166381 -0.39428027  0.06761612  0.067022783
## oldpeak   0.21001257  0.09609288 -0.14923016  0.19321647  0.053951920
## slope    -0.16881424 -0.03071057  0.11971659 -0.12147458 -0.004037770
## ca        0.27632624  0.11826141 -0.18105303  0.10138899  0.070510925
## thal      0.06800138  0.21004110 -0.16173557  0.06220989  0.098802993
## target   -0.22543872 -0.28093658  0.43379826 -0.14493113 -0.085239105
##                 fbs     restecg      thalach       exang      oldpeak
## age       0.121307648 -0.11621090 -0.398521938  0.09680083  0.210012567
## sex       0.045031789 -0.05819627 -0.044019908  0.14166381  0.096092877
## cp        0.094444035  0.04442059  0.295762125 -0.39428027 -0.149230158
## trestbps  0.177530542 -0.11410279 -0.046697728  0.06761612  0.193216472
## chol      0.013293602 -0.15104008 -0.009939839  0.06702278  0.053951920
## fbs       1.000000000 -0.08418905 -0.008567107  0.02566515  0.005747223
```

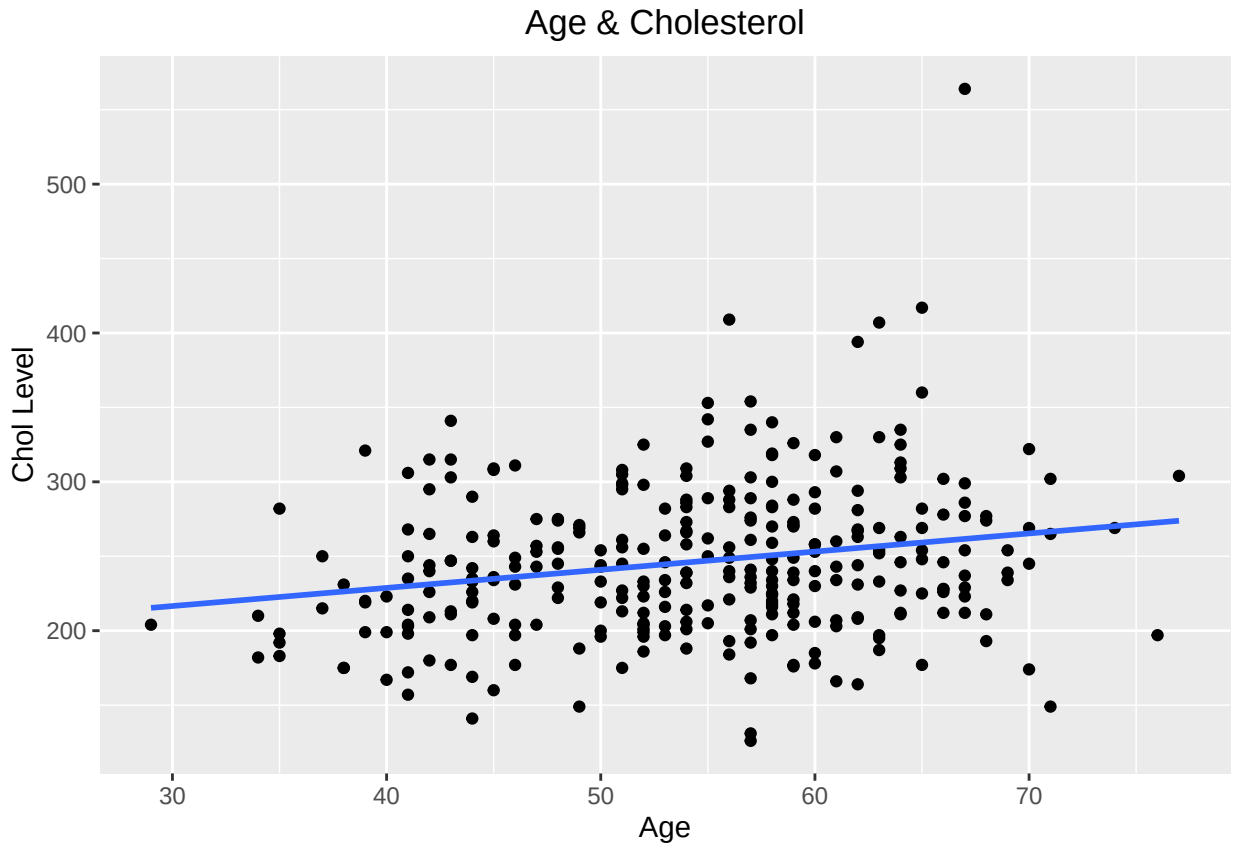
```
## restecg  -0.084189054  1.00000000  0.044123444 -0.07073286 -0.058770226
## thalach  -0.008567107  0.04412344  1.000000000 -0.37881209 -0.344186948
## exang     0.025665147 -0.07073286 -0.378812094  1.00000000  0.288222808
## oldpeak   0.005747223 -0.05877023 -0.344186948  0.28822281  1.000000000
## slope    -0.059894178  0.09304482  0.386784410 -0.25774837 -0.577536817
## ca        0.137979327 -0.07204243 -0.213176928  0.11573938  0.222682322
## thal     -0.032019339 -0.01198140 -0.096439132  0.20675379  0.210244126
## target   -0.028045760  0.13722950  0.421740934 -0.43675708 -0.430696002
##                slope          ca        thal      target
## age      -0.16881424  0.27632624  0.06800138 -0.22543872
## sex      -0.03071057  0.11826141  0.21004110 -0.28093658
## cp        0.11971659 -0.18105303 -0.16173557  0.43379826
## trestbps -0.12147458  0.10138899  0.06220989 -0.14493113
## chol     -0.00403777  0.07051093  0.09880299 -0.08523911
## fbs      -0.05989418  0.13797933 -0.03201934 -0.02804576
## restecg   0.09304482 -0.07204243 -0.01198140  0.13722950
## thalach   0.38678441 -0.21317693 -0.09643913  0.42174093
## exang    -0.25774837  0.11573938  0.20675379 -0.43675708
## oldpeak  -0.57753682  0.22268232  0.21024413 -0.43069600
## slope     1.00000000 -0.08015521 -0.10476379  0.34587708
## ca       -0.08015521  1.00000000  0.15183213 -0.39172399
## thal     -0.10476379  0.15183213  1.00000000 -0.34402927
## target    0.34587708 -0.39172399 -0.34402927  1.00000000
```

```
corrplot(cor_heart, type="upper", order="hclust", tl.col="black", tl.srt=45)
```



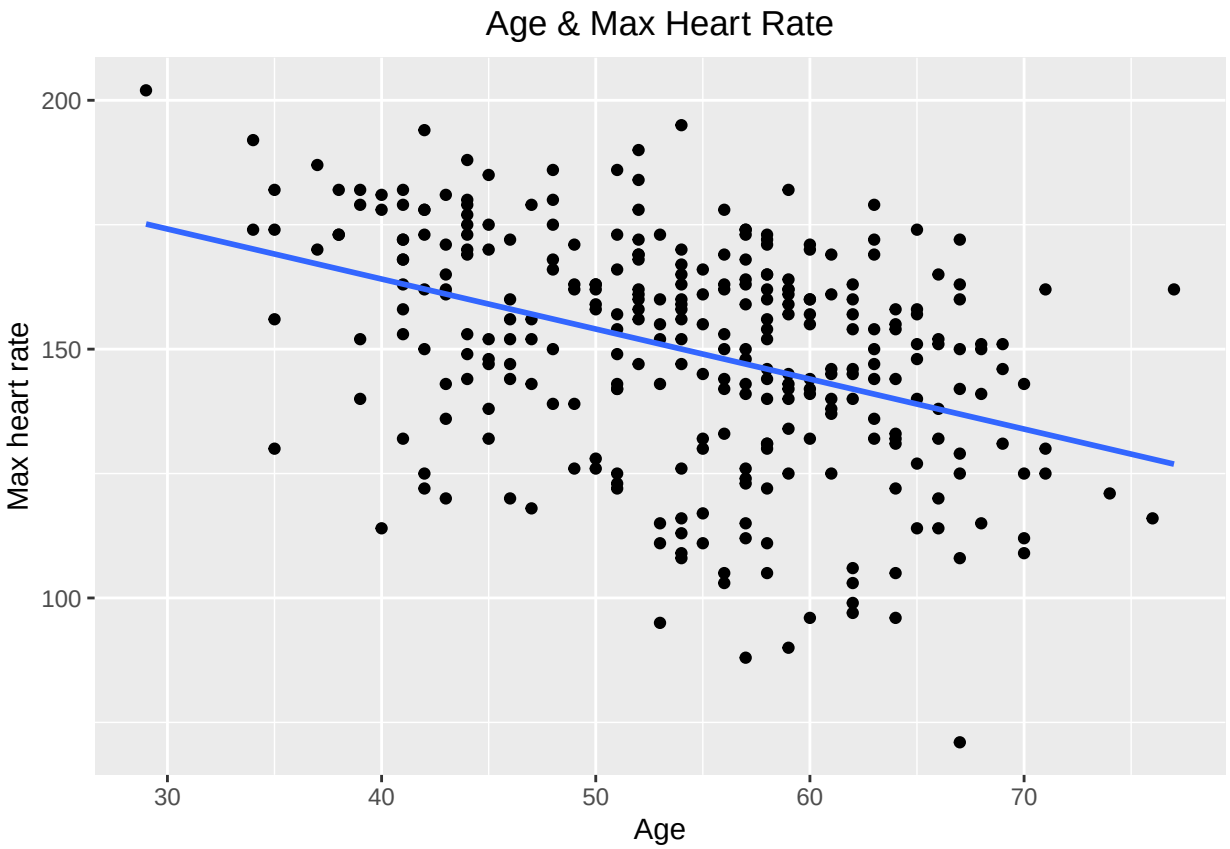
```
g_age_chol <- ggplot(data2,aes(x=age,y=chol))+
    geom_point()+
    geom_smooth(method = "lm", se = FALSE)+
    scale_x_continuous(name="Age")+
    scale_y_continuous(name="Chol Level")+
    ggtitle("Age & Cholesterol")+
    theme(plot.title = element_text(hjust = 0.5))
  g_age_chol
```

```
## `geom_smooth()` using formula 'y ~ x'
```

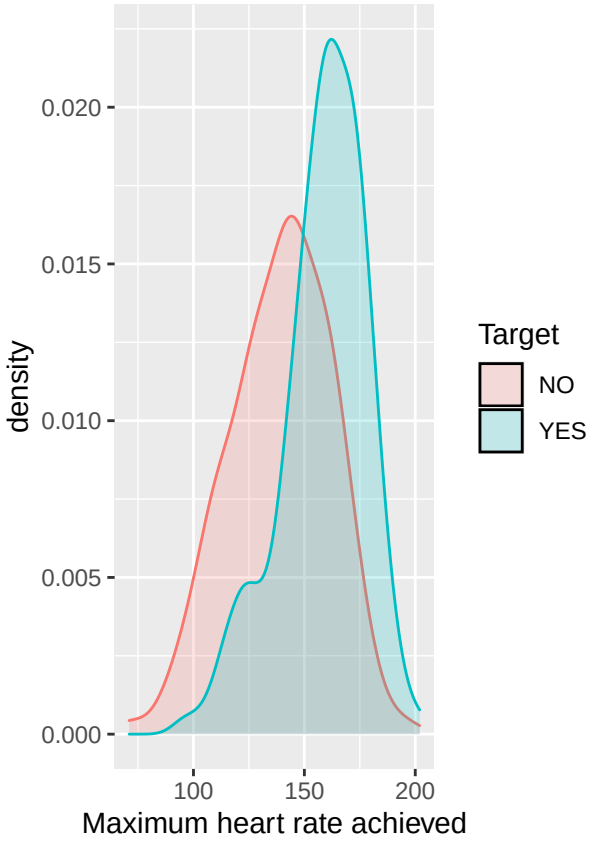


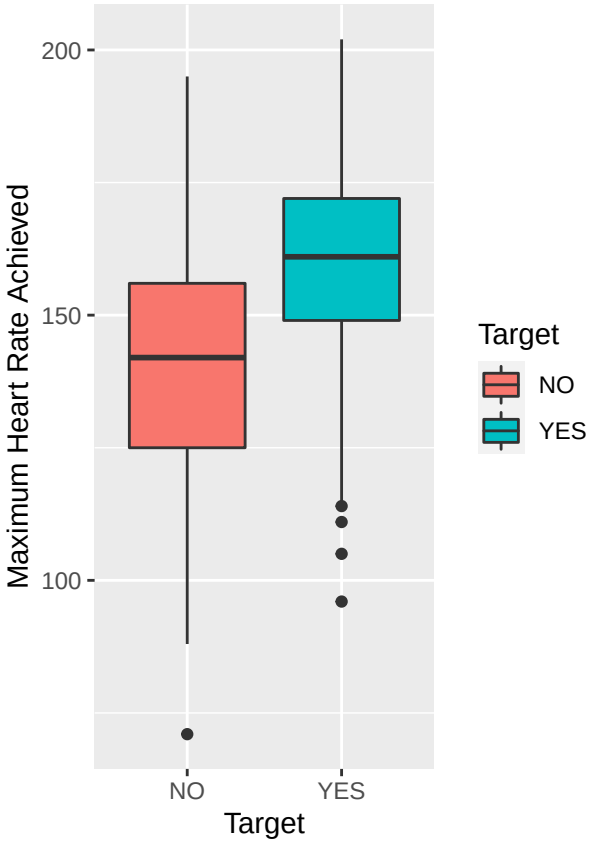

```
  # age and max heart rate
  g_age_maxhr <- ggplot(data2,aes(x=age,y=thalach))+
    geom_point()+geom_smooth(method = "lm", se= FALSE)+
    scale_x_continuous(name="Age")+
    scale_y_continuous(name="Max heart rate")+
    ggtitle("Age & Max Heart Rate")+
    theme(plot.title = element_text(hjust = 0.5))

  g_age_maxhr
```

```
## `geom_smooth()` using formula 'y ~ x'
```

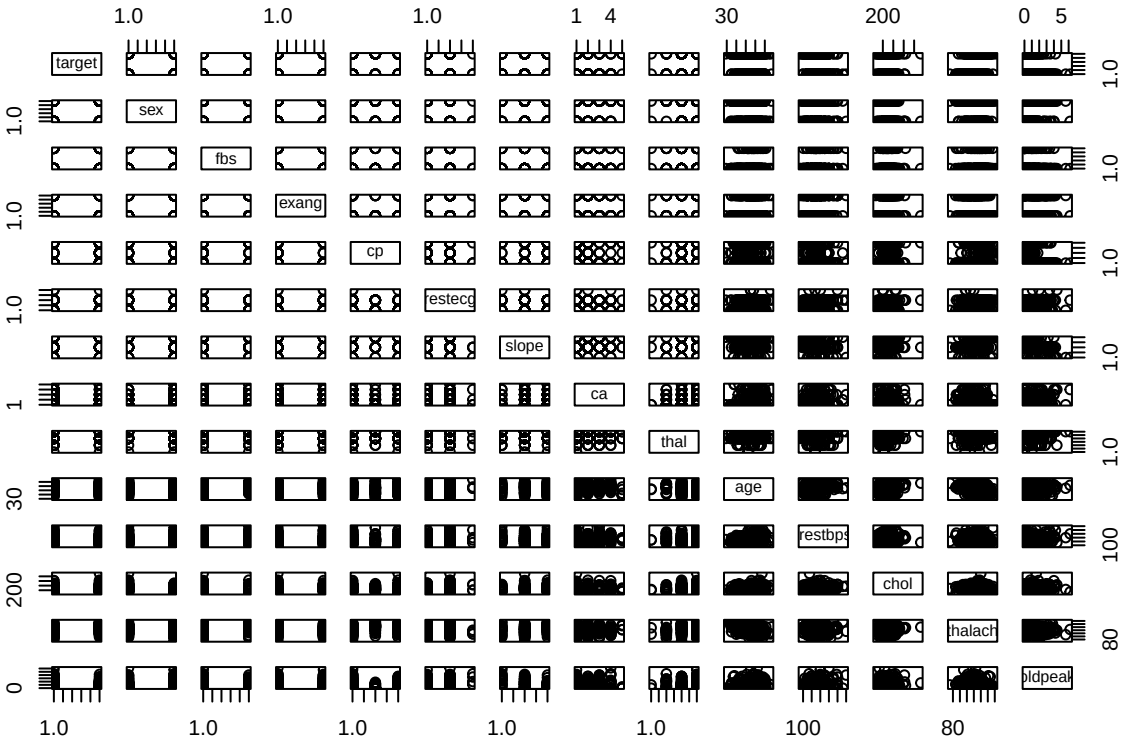



```
g1 <- ggplot(data2,aes(thalach,col=as.factor(target),fill=as.factor(target)))+
    geom_density(alpha=0.2)+
    guides(col="none")+
    labs(fill="Target",x="Maximum heart rate achieved")
    #theme_economist_white(gray_bg = FALSE)

  # max heart rate and target boxplot
g2 <- ggplot(data2,aes(as.factor(target),thalach,fill=as.factor(target)))+
    geom_boxplot()+
    labs(y="Maximum Heart Rate Achieved",x="Target",fill="Target")
    #theme_economist_white(gray_bg = FALSE)


grid.arrange(g1, g2, nrow = 1)
```


0.020
0.015
0.010
0.005
0.000
density
100
150
200
Maximum heart rate achieved
Target
NO
YES


200
150
100
Maximum Heart Rate Achieved
NO
YES
Target
Target
NO
YES

```
pairs(data2)
```


target
sex
fbs
exang
cp
restecg
slope
ca
thal
age
restbps
chol
thalach
oldpeak