Excercise 1

Aaron Green

```r
## Loading necessary packages
library(ggplot2) # For Data visualisation
library(tidyverse) # For data Manipulation
```

```
-- Attaching core tidyverse packages ---------------------- tidyverse 2.0.0 --
v dplyr     1.1.4      v readr     2.1.5
v forcats   1.0.0      v stringr   1.5.1
v lubridate 1.9.4      v tibble    3.2.1
v purrr     1.0.2      v tidyr     1.3.1
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```r
# Fetching the Heart disease data from the given link
link <- "https://www.statlearning.com/s/Heart.csv"
heart_disease <- read.csv(link)

# Exploring the data
head(heart_disease)
```

```
  X Age Sex      ChestPain RestBP Chol Fbs RestECG MaxHR ExAng Oldpeak Slope Ca
1 1  63   1        typical    145  233   1       2   150     0     2.3     3  0
2 2  67   1   asymptomatic    160  286   0       2   108     1     1.5     2  3
3 3  67   1   asymptomatic    120  229   0       2   129     1     2.6     2  2
4 4  37   1      nonanginal    130  250   0       0   187     0     3.5     3  0
5 5  41   0     nontypical    130  204   0       2   172     0     1.4     1  0
6 6  56   1     nontypical    120  236   0       0   178     0     0.8     1  0
       Thal AHD
1     fixed  No
2    normal Yes
```

```
3 reversable Yes
4     normal  No
5     normal  No
6     normal  No


################################################################
# DATA MANIPULATION & TRANSFROMATION
################################################################

# Removing the first column in the dataset as it is just row numbers
heart_disease <- heart_disease[, -1]

# Converting variable names to lowercase to simplify data handling
names(heart_disease) <- tolower(names(heart_disease))

# explicitly define the variables chestpain, thal and ahd to match the specified levels

heart_disease <- heart_disease %>%
  mutate(chestpain = recode(chestpain,
                      "asymptomatic" = 0,
                      "nontypical" = 1,
                      "nonanginal" = 2,
                      "typical" = 3),
         thal = recode(thal,
                   "fixed" = 1,
                   "normal" = 2,
                   "reversable" = 3),
         ahd = recode(ahd,
                   "No" = 1,
                   "Yes" = 0))

# renaming ahd to target for better understanding
heart_disease <- heart_disease %>%
  rename(target = ahd)

# FACTORING VARIABLES ACCORDINGLY
#NB: slope has been coded with 1, 2, 3 to match content in the dataset
heart_disease <- heart_disease %>%
  mutate(sex = factor(sex,
                   levels = c(0, 1),
                   labels = c("female", "male")),
         chestpain = factor(chestpain,
```

```r
                        levels = c(0, 1, 2, 3),
                        labels = c("asymptomatic", "nontypical",
                                   "nonanginal", "typical")),
        fbs = factor(fbs,
                     levels = c(0, 1),
                     labels = c("false", "true")),
        restecg = factor(restecg,
                         levels = c(0, 1, 2),
                         labels = c("showing probable or definite left ventricular hypertrop
                                    "normal",
                                    "having ST-T wave abnormality")),
        exang = factor(exang,
                       levels = c(0, 1),
                       labels = c("no", "yes")),
        slope = factor(slope,
                       levels = c(1, 2, 3),
                       labels = c("downsloping", "flat", "upsloping")),
        thal = factor(thal,
                      levels = c(1, 2, 3),
                      labels = c("fixed", "normal", "reversable")),
        target = factor(target,
                        levels = c(0, 1),
                        labels = c("yes", "no"))

        )
str(heart_disease)
```

```
'data.frame':   303 obs. of  14 variables:
 $ age      : int  63 67 67 37 41 56 62 57 63 53 ...
 $ sex      : Factor w/ 2 levels "female","male": 2 2 2 2 1 2 1 1 2 2 ...
 $ chestpain: Factor w/ 4 levels "asymptomatic",..: 4 1 1 3 2 2 1 1 1 1 ...
 $ restbp   : int  145 160 120 130 130 120 140 120 130 140 ...
 $ chol     : int  233 286 229 250 204 236 268 354 254 203 ...
 $ fbs      : Factor w/ 2 levels "false","true": 2 1 1 1 1 1 1 1 1 2 ...
 $ restecg  : Factor w/ 3 levels "showing probable or definite left ventricular hypertrophy l
 $ maxhr    : int  150 108 129 187 172 178 160 163 147 155 ...
 $ exang    : Factor w/ 2 levels "no","yes": 1 2 2 1 1 1 1 1 2 1 2 ...
 $ oldpeak  : num  2.3 1.5 2.6 3.5 1.4 0.8 3.6 0.6 1.4 3.1 ...
 $ slope    : Factor w/ 3 levels "downsloping",..: 3 2 2 3 1 1 3 1 2 3 ...
 $ ca       : int  0 3 2 0 0 0 2 0 1 0 ...
 $ thal     : Factor w/ 3 levels "fixed","normal",..: 1 2 3 2 2 2 2 2 3 3 ...
 $ target   : Factor w/ 2 levels "yes","no": 2 1 1 2 2 2 1 2 1 1 ...
```

```
################################################################
# PLOTS
################################################################

# Plot 1: Distribution of heart disease by sex

heart_disease %>%
  ggplot(aes(x = sex,
             fill = target)) +
  geom_bar(position = "dodge") +
  geom_text(stat = "count",
            aes(label = after_stat(count), # show bar counts as labels
                color = target),           # color labels by target to match the fill colour
            position = position_dodge(width = 0.8), # align labels with bars
            vjust = -0.28,  # nudge labels above bars
            size = 3.2) +  # set label font size
  scale_fill_brewer(palette = "Set1") + # apply fill palette
  scale_color_brewer(palette = "Set1", guide = "none") + # match label colors to fill colour
  labs(title = "Distribution of heart disease by sex",
       x = "",
       y = "Count",
       caption = "n = 303",
       fill = "Heart disease") +
  ggthemes::theme_clean()
```

Distribution of heart disease by sex
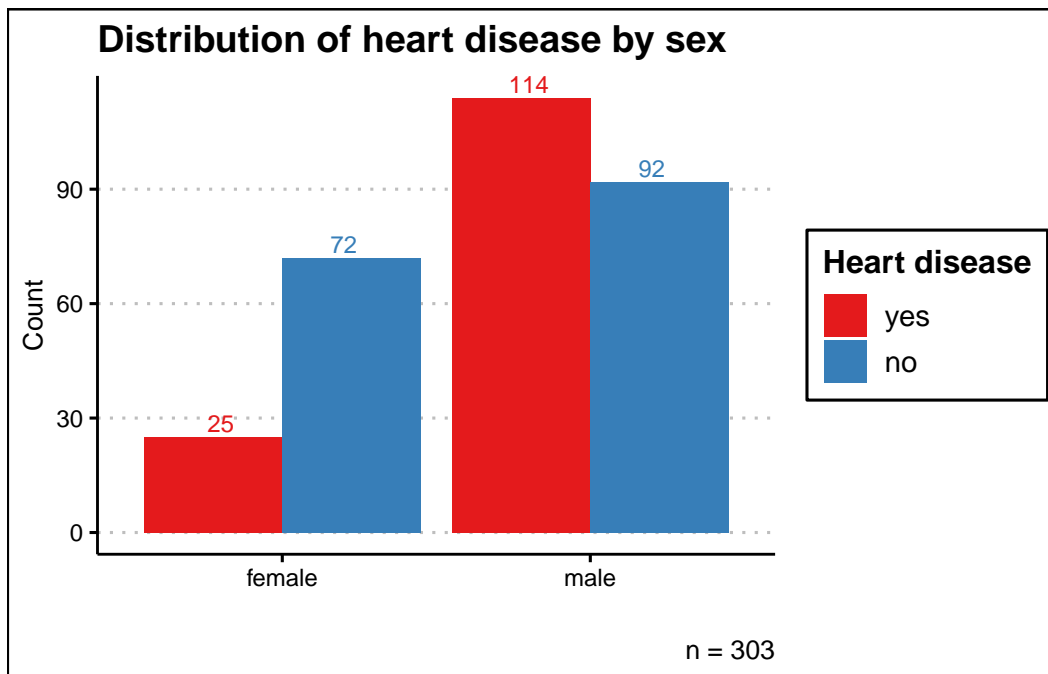

```
# Plot 2: Excercise induced angina by sex
heart_disease %>%
  filter(exang == "yes") %>%
  ggplot(aes(x = sex,
             fill = sex)) +
  geom_bar(position = "dodge") +
  geom_text(stat = "count",
            aes(label = after_stat(count)),
            vjust = 1.5,           # place inside the bar
            color = "white",       # white text for contrast
            fontface = "bold",     # bold text
            size = 4) +            # font size
  scale_fill_brewer(palette = "Set1") +
  labs(title = "Excercise induced angina by sex",
       x = "",
       y = "Count",
       caption = "n = 99", ) +
  ggthemes::theme_clean() +
    theme(legend.position = "none") # removing the legend as it is not necessary
```

Excercise induced angina by sex

Count

80
77
60
40
22
20
0

female   male

n = 99