## SDS 395 Report

*Sookja Kang, sk26949*

### Report

```
HA <- read_csv("heart.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   age = col_double(),
##   sex = col_double(),
##   cp = col_double(),
##   trtbps = col_double(),
##   chol = col_double(),
##   fbs = col_double(),
##   restecg = col_double(),
##   thalachh = col_double(),
##   exng = col_double(),
##   oldpeak = col_double(),
##   slp = col_double(),
##   caa = col_double(),
##   thall = col_double(),
##   output = col_double()
## )
```

```
summary(HA)
```

```
##       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
```

#Question: 1) Are there age differences for people who had exercise induced angina or did not across male and female groups? 2) How do the proportion of different chest pain types change across the gender groups?

**Introduction:**

*I am using `heart.csv` to answer Question 1. This dataset contains 303 individuals with 14 categories of their health information related to the heart. Heart related health information includes age, sex, chest pain type(cp), resting blood pressure(trtbps), cholesterol level(chol), fasting blood sugar level(fbs), resting electrocardiographic result (restecg), maximum heart rate(thalachh), previous peak(oldpeak), slop(slp), number of major vessels(caa), Thalium Stress Test result(thall), exercise-induced angina (exng), and heart attack or not (output). To understand the relationships of age, chest pain types, and angina occurrence across the gender groups, I am going to use the four variables: age, sex, chest pain type, and exercise-induced angina.*

1. age: a numeric variable
2. sex: reported 0 as female and 1 as Male
3. chest pain type: reported 0 as Typical Angina, 1 as Atypical Angina, 2 as Non-anginal Pain, and 3 as Asymptomatic
4. exercise induced angina: reported 0 as no and 1 as yes

**Approach:**

*My approach is to understand 1) the age differences of people with or without exercise-induced angina and 2) the proportion of different chest pain types change across the gender groups. First, I am going to make violin plots to show age distributions of people with or without exercise-induced angina. These violin plots will allow comparing the age distributions across the different groups side by side. Next, I am going to make a pie chart to visually show the proportion changes of chest pain types across the gender group. The pie charts will allow me to easily compare each slice's proportion (different chest pain types) in a whole circle across the different gender groups.*

To make violin plots, the following functions will be used: 1. factor(): to encode a vector as a factor 2. geom_violin(): to make a vilon plot using the data

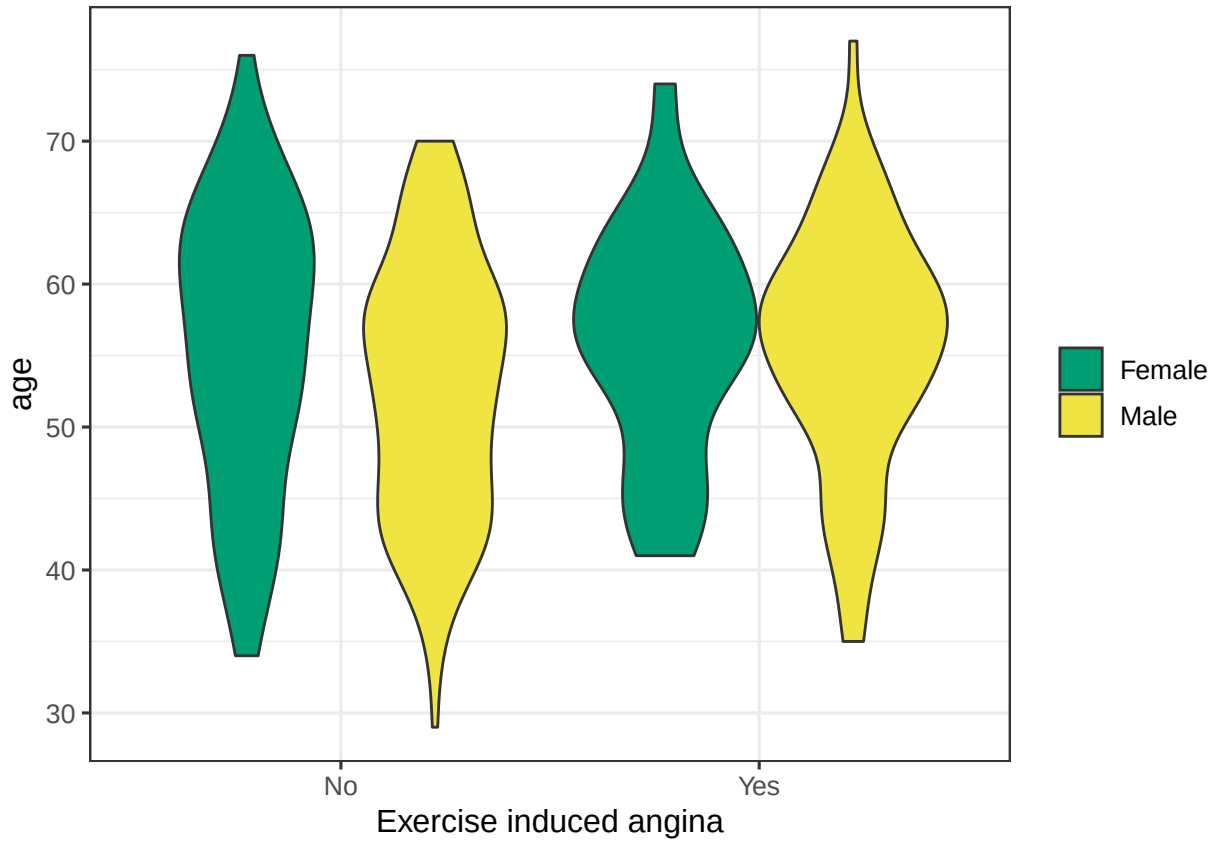
To make pie charts, the following functions will be used: 1. factor(): to encode a vector as a factor 2. labels(): to label the values 3. count(): count numbers for each subcategory of chest pain type and sex 4. mutate(): make a new column (total_number) using the n column that created from count() 5. arrange() and -desc(): to sort the total_number by ascending count 6. fct_reorder(): to reorder the chest pain type column by the total_number 7. group_by(): to group by the sex 8. mutate(): make new columns • the end_angle, start_angle, mid_angle for each pie slice • horizontal and vertical justifications for outer labels 9. ggplot(): to plot the pie_data 10. geom_arc_bar() to specify the exact location of the pie center in the x-y plane 11. coord_fixed(): to ensure that the pie is round 12. facet_wrap(): to create pie chart facets for each income level 13. theme_void(): to remove the x-y plane

**Analysis:**

*Violin Plot*

```
ggplot(HA,
       aes(x = factor(exng), y = age, fill = factor(sex))
       ) +
  geom_violin() +
  scale_x_discrete(
    name = "Exercise induced angina",
    labels = c ("No", "Yes")
    ) +
   scale_fill_manual(
```

```
    name = NULL,
    labels = c("Female", "Male"),
    values = c("#009E73", "#F0E442")
    ) +
  theme_bw(12)
```




*Pie Chart*

```
HA$cp <- factor(HA$cp,
                levels = c(0, 1, 2, 3),
                labels = c("Typical Angina", "Atypical Angina", "Non-anginal Pain", "Asymptomatic"))
HA$sex <- factor(HA$sex,
                 levels = c(0, 1),
                 labels = c("Female", "Male"))


HA_data <- HA %>%
  count(cp, sex) %>%
  mutate(total_number = n ) %>%
  arrange(-desc(total_number))  %>%
  mutate(cp = fct_reorder(cp, total_number))

HA_data
```

```
## # A tibble: 8 x 4
##   cp               sex        n total_number
##   <fct>            <fct>  <int>        <int>
```

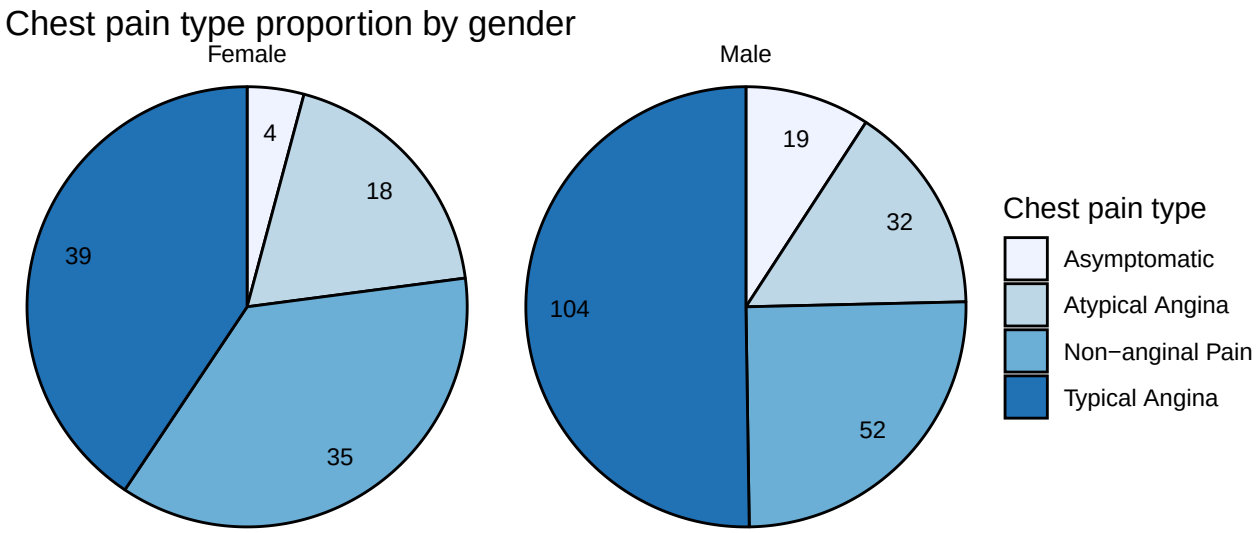
```
## 1 Asymptomatic     Female     4            4
## 2 Atypical Angina  Female    18           18
## 3 Asymptomatic     Male      19           19
## 4 Atypical Angina  Male      32           32
## 5 Non-anginal Pain Female    35           35
## 6 Typical Angina   Female    39           39
## 7 Non-anginal Pain Male      52           52
## 8 Typical Angina   Male     104          104
```

```
pie_data<- HA_data %>%
  group_by(sex) %>%
  mutate(end_angle = 2*pi*cumsum(n)/sum(n),
         start_angle = lag(end_angle, default = 0),
         mid_angle = 0.5*(start_angle + end_angle),
         hjust = ifelse(mid_angle > pi, 1, 0),
         vjust = ifelse(mid_angle < pi/2 | mid_angle > 3*pi/2, 0, 1))

pie_data
```

```
## # A tibble: 8 x 9
## # Groups:   sex [2]
##   cp        sex       n total_number end_angle start_angle mid_angle hjust vjust
##   <fct>     <fct> <int>        <int>     <dbl>       <dbl>     <dbl> <dbl> <dbl>
## 1 Asymptom~ Fema~     4            4     0.262       0         0.131     0     0
## 2 Atypical~ Fema~    18           18     1.44        0.262     0.851     0     0
## 3 Asymptom~ Male     19           19     0.577       0         0.288     0     0
## 4 Atypical~ Male     32           32     1.55        0.577     1.06      0     0
## 5 Non-angi~ Fema~    35           35     3.73        1.44      2.59      0     1
## 6 Typical ~ Fema~    39           39     6.28        3.73      5.01      1     0
## 7 Non-angi~ Male     52           52     3.13        1.55      2.34      0     1
## 8 Typical ~ Male    104          104     6.28        3.13      4.70      1     1
```

```
ggplot(pie_data, aes(x0 = 0, y0 = 0, r0 = 0, r = 1,
                     start = start_angle, end = end_angle,
                     fill = cp)
       ) +
  geom_arc_bar() +
  geom_text(size = 3,
            aes(x = 0.8 * sin(mid_angle),
                y = 0.8 * cos(mid_angle),
                label = total_number)
            ) +
  coord_fixed() +
  facet_wrap(~sex) +
  theme_void() +
  scale_fill_brewer(name = "Chest pain type") +
  ggtitle("Chest pain type proportion by gender")
```




**Discussion:**

*The violin plots show age distributions of people with or without exercise-induced angina across the gender groups. The groups with no exercise-induced angina show violin plots starting at a younger age compared to the exercise-induced angina group. The groups with exercise-induced angina show violin plots ending at a similar or older age compared to no exercise-induced angina group. When I compare the gender groups in the exercised group, the male group has wider age distribution than the female group (both groups show similar median points). Males start experiencing exercise-induced angina compared to females.*

*The pie charts show how the proportions of chest pain types change across the two gender groups. Both groups show similar patterns of chest pain types: Typical angina > Non-anginal pain > Atypical angina > Asymptomatic. The female group shows a higher proportion of Non-anginal pain and a lower proportion of asymptomatic compared to the male group.*