# Midterm 2 W25

Javier Vidal

2025-03-04

# Instructions

Before starting the exam, you need to follow the instructions in `02_midterm2_cleaning.Rmd` to clean the data. Once you have cleaned the data and produced the `heart.csv` file, you can start the exam.

Answer the following questions and complete the exercises in RMarkdown. Please embed all of your code and push your final work to your repository. Your code must be organized, clean, and run free from errors. Remember, you must remove the `#` for any included code chunks to run. Be sure to add your name to the author header above.

Your code must knit in order to be considered. If you are stuck and cannot answer a question, then comment out your code and knit the document. You may use your notes, labs, and homework to help you complete this exam. Do not use any other resources- including AI assistance or other students' work.

Don't forget to answer any questions that are asked in the prompt! Each question must be coded; it cannot be answered by a sort in a spreadsheet or a written response.

All plots should be clean, with appropriate labels, and consistent aesthetics. Poorly labeled or messy plots will receive a penalty. Your plots should be in color and look professional!

Be sure to push your completed midterm to your repository and upload the document to Gradescope. This exam is worth 30 points.

# Load the libraries

You may not use all of these, but they are here for convenience.

```
library("tidyverse")
library("janitor")
library("ggthemes")
library("RColorBrewer")
library("paletteer")
```

# Load the data

These data are a modified version of the Statlog (Heart) database on heart disease from the UCI Machine Learning Repository (https://archive.ics.uci.edu/dataset/145/statlog+heart). The data are also available on Kaggle (https://www.kaggle.com/datasets/ritwikb3/heart-disease-statlog/data).

You will need the descriptions of the variables to answer the questions. Please reference `03_midterm2_descriptions.Rmd` for details.

Run the following to load the data.

```
heart <- read_csv("data/heart.csv")
```

# Questions

Problem 1. (1 point) Use the function of your choice to provide a data summary.

```
glimpse(heart)
```

```
## Rows: 270
## Columns: 14
## $ age      <dbl> 70, 67, 57, 64, 74, 65, 56, 59, 60, 63, 59, 53, 44, 61, 57, 7…
## $ gender   <chr> "male", "female", "male", "male", "female", "male", "male", "…
## $ cp       <chr> "asymptomatic", "non_anginal_pain", "atypical_angina", "asymp…
## $ trestbps <dbl> 130, 115, 124, 128, 120, 120, 130, 110, 140, 150, 135, 142, 1…
## $ chol     <dbl> 322, 564, 261, 263, 269, 177, 256, 239, 293, 407, 234, 226, 2…
## $ fbs      <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, TRUE, FALSE, FALSE,…
## $ restecg  <chr> "left_ventricular_hypertrophy", "left_ventricular_hypertrophy…
## $ thalach  <dbl> 109, 160, 141, 105, 121, 140, 142, 142, 170, 154, 161, 111, 1…
## $ exang    <chr> "no", "no", "no", "yes", "yes", "no", "yes", "yes", "no", "no…
## $ oldpeak  <dbl> 2.4, 1.6, 0.3, 0.2, 0.2, 0.4, 0.6, 1.2, 1.2, 4.0, 0.5, 0.0, 0…
## $ slope    <chr> "flat", "flat", "upsloping", "flat", "upsloping", "upsloping"…
## $ ca       <dbl> 3, 0, 0, 1, 1, 0, 1, 1, 2, 3, 0, 0, 0, 2, 1, 0, 2, 0, 0, 0, 2…
## $ thal     <chr> "normal", "reversable_defect", "reversable_defect", "reversab…
## $ target   <chr> "disease", "no_disease", "disease", "no_disease", "no_disease…
```

Problem 2. (1 point) Let's explore the demographics of participants included in the study. What is the number of males and females? Show this as a table.

```
heart %>%
  count(gender)
```

```
## # A tibble: 2 × 2
##   gender     n
##   <chr>  <int>
## 1 female    87
## 2 male     183
```

Problem 3. (2 points) What is the average age of participants by gender? Show this as a table.

```
heart %>%
  group_by(gender) %>%
  summarize(average_age = mean(age))
```

```
## # A tibble: 2 × 2
##   gender average_age
##   <chr>        <dbl>
## 1 female        55.7
## 2 male          53.8
```

Average age for female is 55 and average age for male is 53

Problem 4. (1 point) Among males and females, how many have/do not have heart disease? Show this as a table, grouped by gender.

```
heart %>%
  group_by(gender) %>%
  count(target)
```

```
## # A tibble: 4 × 3
## # Groups:   gender [2]
##   gender target         n
##   <chr>  <chr>      <int>
## 1 female disease       20
## 2 female no_disease    67
## 3 male   disease      100
## 4 male   no_disease    83
```

20 females have disease, 67 do not. 100 males have disease, 83 do not

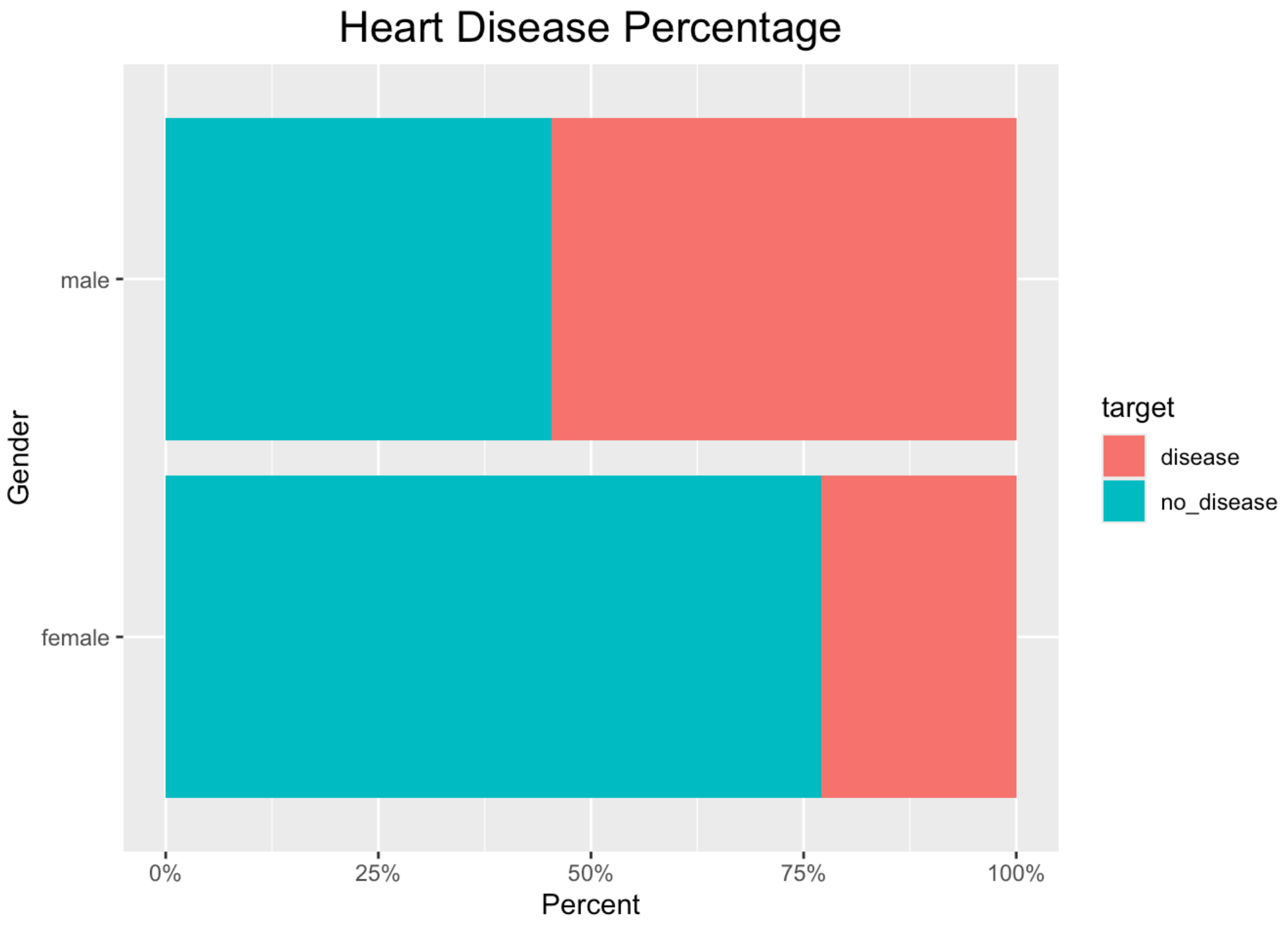
Problem 5. (4 points) What is the percentage of males and females with heart disease? Show this as a table, grouped by gender.

```
heart %>%
  filter(target == "disease") %>%
  tabyl(gender)
```

```
##   gender   n   percent
##   female  20 0.1666667
##     male 100 0.8333333
```

Problem 6. (3 points) Make a plot that shows the results of your analysis from problem 5. If you couldn’t get the percentages to work, then make a plot that shows the number of participants with and without heart disease by gender.

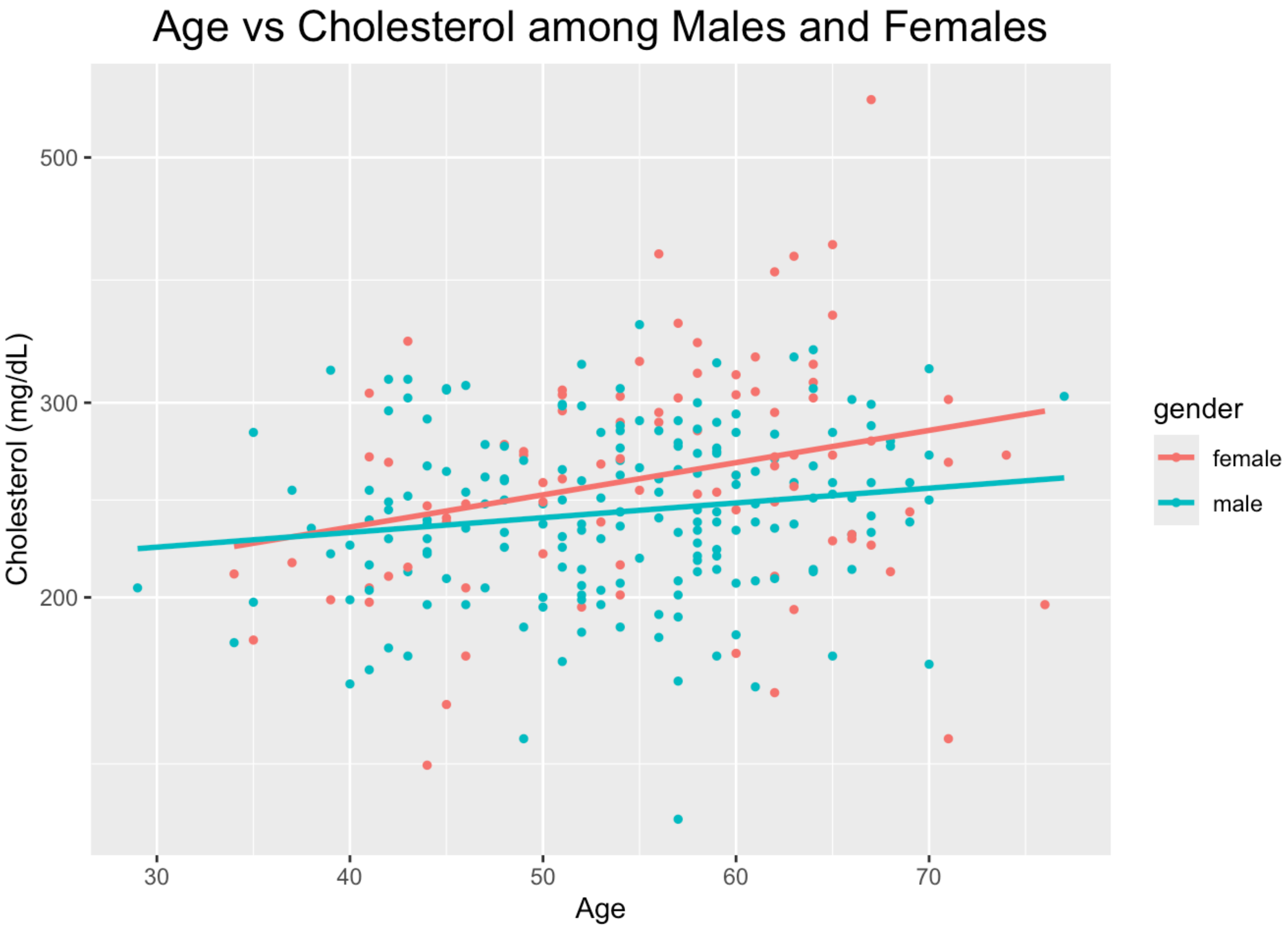
```
heart %>%
  ggplot(aes(x=gender, fill=target))+
  geom_bar(position=position_fill())+
  scale_y_continuous(labels=scales::percent)+
  coord_flip()+
  theme_grey()+
  labs(title = "Heart Disease Percentage",
       x="Gender",
       y="Percent")+
  theme(plot.title=element_text(size=rel(1.5), hjust=.5))
```



Problem 7. (3 points) Is there a relationship between age and cholesterol levels? Make a plot that shows this relationship separated by gender (hint: use faceting or make two plots). Be sure to add a line of best fit (linear regression line).

```
heart %>%
  ggplot(aes(x = age, y = chol, color = gender))+
  geom_point(size = 1)+
  geom_smooth(method = lm, se = F)+
  scale_y_log10()+
  theme_grey()+
  labs(title = "Age vs Cholesterol among Males and Females",
       x="Age",
       y="Cholesterol (mg/dL)")+
  theme(plot.title=element_text(size=rel(1.5), hjust=.5))
```

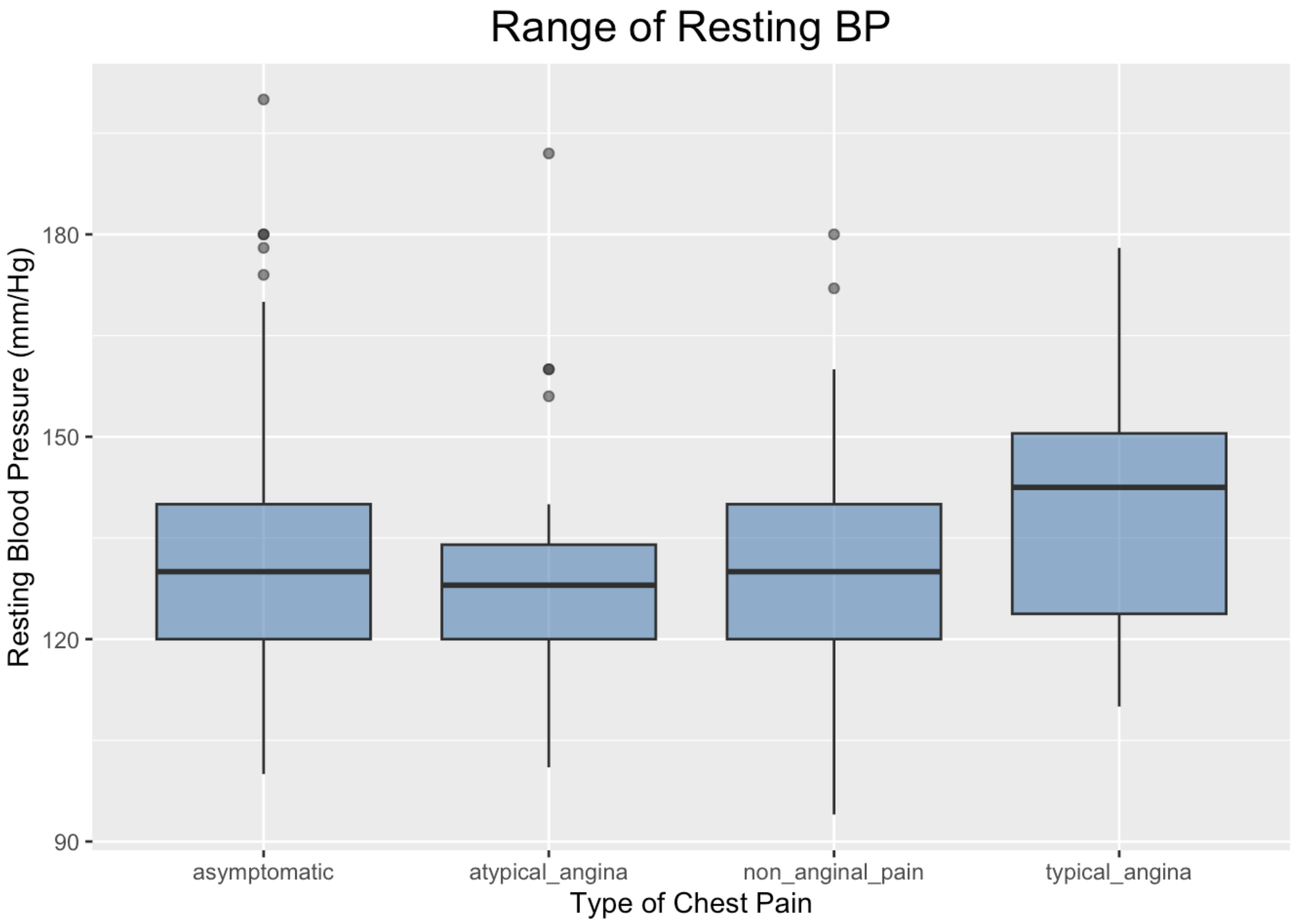
```
## `geom_smooth()` using formula = 'y ~ x'
```



There is a relationship between age and cholesterol levels. Cholesterol levels tend to increase with age.

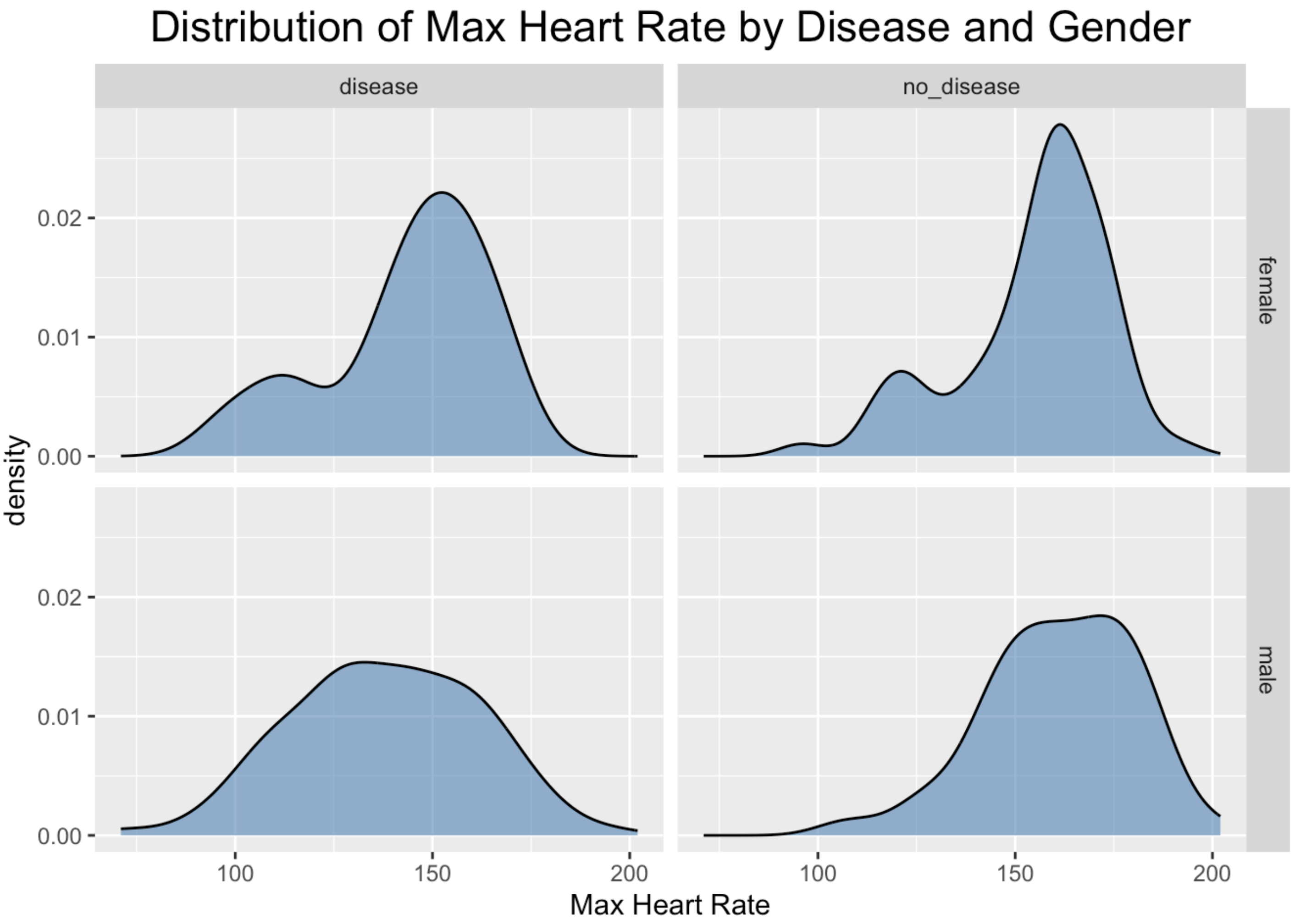
Problem 8. (3 points) What is the range of resting blood pressure for participants by type of chest pain? Make a plot that shows this information.

```
heart %>%
  ggplot(aes(x= cp, y = trestbps))+
  geom_boxplot(fill = "steelblue", alpha = 0.6)+
  theme_grey()+
  labs(title = "Range of Resting BP",
       x="Type of Chest Pain",
       y="Resting Blood Pressure (mm/Hg)")+
  theme(plot.title=element_text(size=rel(1.5), hjust=.5))
```



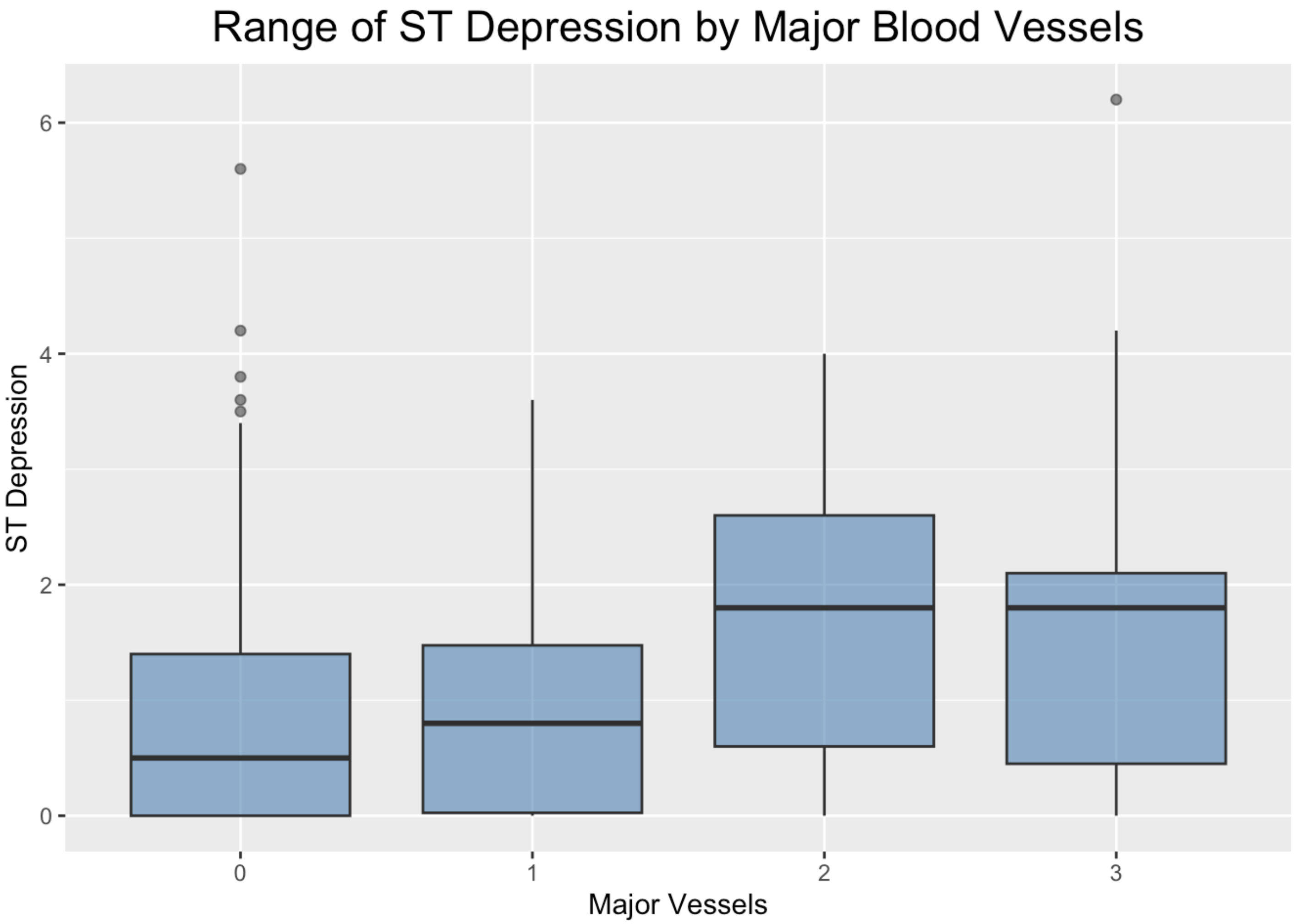
Problem 9. (4 points) What is the distribution of maximum heart rate achieved, separated by gender and whether or not the patient has heart disease? Make a plot that shows this information- you must use faceting.

```
heart %>%
  ggplot(aes(x=thalach))+
  geom_density(fill = "steelblue", alpha = 0.6)+
  facet_grid(gender~target)+
  theme_grey()+
  labs(title = "Distribution of Max Heart Rate by Disease and Gender",
       x="Max Heart Rate")+
  theme(plot.title=element_text(size=rel(1.5), hjust=.5))
```



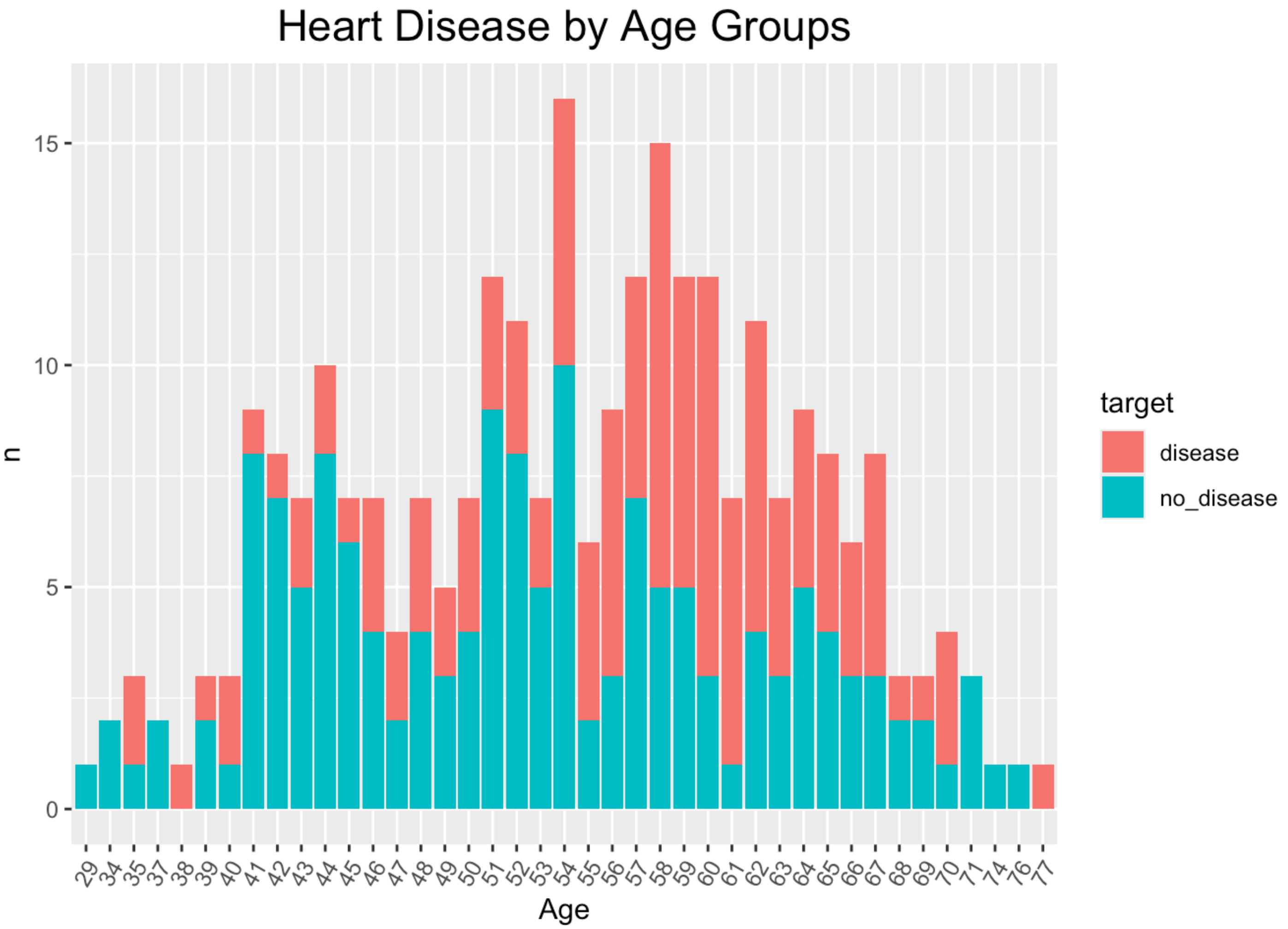
Problem 10. (4 points) What is the range of ST depression (oldpeak) by the number of major vessels colored by fluoroscopy (ca)? Make a plot that shows this relationship. (hint: should ca be a factor or numeric variable?)

```
heart %>%
  ggplot(aes(x = as_factor(ca), y = oldpeak))+
  geom_boxplot(fill = "steelblue", alpha = 0.6)+
  theme_grey()+
  labs(title = "Range of ST Depression by Major Blood Vessels",
       x="Major Vessels",
       y="ST Depression")+
  theme(plot.title=element_text(size=rel(1.5), hjust=.5))
```



Problem 11. (4 points) Is there an age group where we see increased prevalence of heart disease? Make a plot that shows the number of participants with and without heart disease by age group.

```
heart %>%
  ggplot(aes(x= as_factor(age), fill=target))+
  geom_bar()+
  theme_grey()+
  labs(title = "Heart Disease by Age Groups",
       x= "Age",
       y="n")+
  theme(plot.title=element_text(size=rel(1.5), hjust=.5))+
   theme(axis.text.x=element_text(angle=60, hjust=1))
```



At age 58, there is an increaased prevalence in heart disease compared to those without heart disease at age 58.

# Submit the Midterm

1. Save your work and knit the .rmd file.
2. Open the .html file and “print” it to a .pdf file in Google Chrome (not Safari).
3. Go to the class Canvas page and open Gradescope.
4. Submit your .pdf file to the midterm assignment- be sure to assign the pages to the correct questions.
5. Commit and push your work to your repository.