# Midterm 2 W25

**Joyce Lin**

**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")
```

```
colors <- paletteer::palettes_d_names
my_palette <- paletteer_d("ochRe::mccrea")
```

# 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…
```

```
summary(heart)
```

```
##       age           gender               cp              trestbps
##  Min.   :29.00   Length:270         Length:270         Min.   : 94.0
##  1st Qu.:48.00   Class :character   Class :character   1st Qu.:120.0
##  Median :55.00   Mode  :character   Mode  :character   Median :130.0
##  Mean   :54.43                                         Mean   :131.3
##  3rd Qu.:61.00                                         3rd Qu.:140.0
##  Max.   :77.00                                         Max.   :200.0
##       chol          fbs            restecg            thalach
##  Min.   :126.0   Mode :logical   Length:270         Min.   : 71.0
##  1st Qu.:213.0   FALSE:230       Class :character   1st Qu.:133.0
##  Median :245.0   TRUE :40        Mode  :character   Median :153.5
##  Mean   :249.7                                      Mean   :149.7
##  3rd Qu.:280.0                                      3rd Qu.:166.0
##  Max.   :564.0                                      Max.   :202.0
##     exang             oldpeak        slope                ca
##  Length:270         Min.   :0.00   Length:270         Min.   :0.0000
##  Class :character   1st Qu.:0.00   Class :character   1st Qu.:0.0000
##  Mode  :character   Median :0.80   Mode  :character   Median :0.0000
##                     Mean   :1.05                      Mean   :0.6704
##                     3rd Qu.:1.60                      3rd Qu.:1.0000
##                     Max.   :6.20                      Max.   :3.0000
##      thal              target
##  Length:270         Length:270
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

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.
*87 females and 183 males*

```
heart %>%
  count(gender) %>%
  arrange(-n)
```

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

Problem 3. (2 points) What is the average age of participants by gender? Show this as a table. *Female has an average age of 55.67816 and male has an average of of 53.84153*

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

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

Problem 4. (1 point) Among males and females, how many have/do not have heart disease? Show this as a table, grouped by gender.
*There are 100 males that have heart disease and 83 males that have no heart disease. For female participants, there are 20 have heart disease and 67 that have no heart disease.*

```
heart %>%
  group_by(gender) %>%
  count(target) %>%
  arrange(-n)
```

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

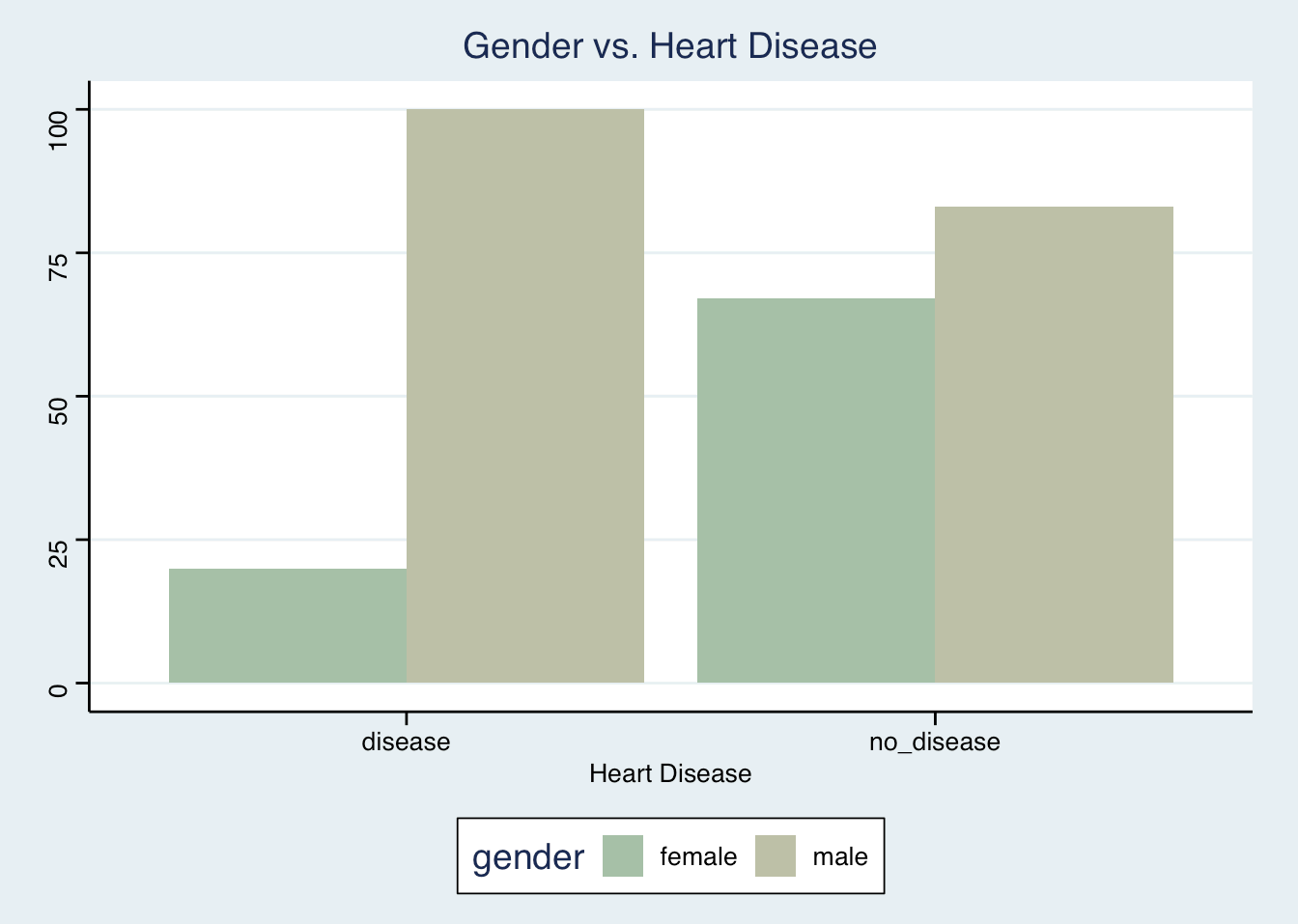
Problem 5. (4 points) What is the percentage of males and females with heart disease? Show this as a table, grouped by gender.
*37.04% Male with disease, 30.74% male with no disease, 7.41% female with disease, and 24.81% female with no disease*

```
heart %>%
  group_by(gender) %>%
  count(target) %>%
  mutate(pct_n=n/270*100) %>%
  arrange(-pct_n)
```

```
## # A tibble: 4 × 4
## # Groups:   gender [2]
##   gender target         n pct_n
##   <chr>  <chr>      <int> <dbl>
## 1 male   disease      100 37.0
## 2 male   no_disease    83 30.7
## 3 female no_disease    67 24.8
## 4 female disease       20  7.41
```

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=target, fill=gender))+
  geom_bar(position = "dodge")+
  labs(title = "Gender vs. Heart Disease",
       x="Heart Disease",
       y=NULL)+
  theme_stata()+
  scale_fill_manual(values=my_palette)
```



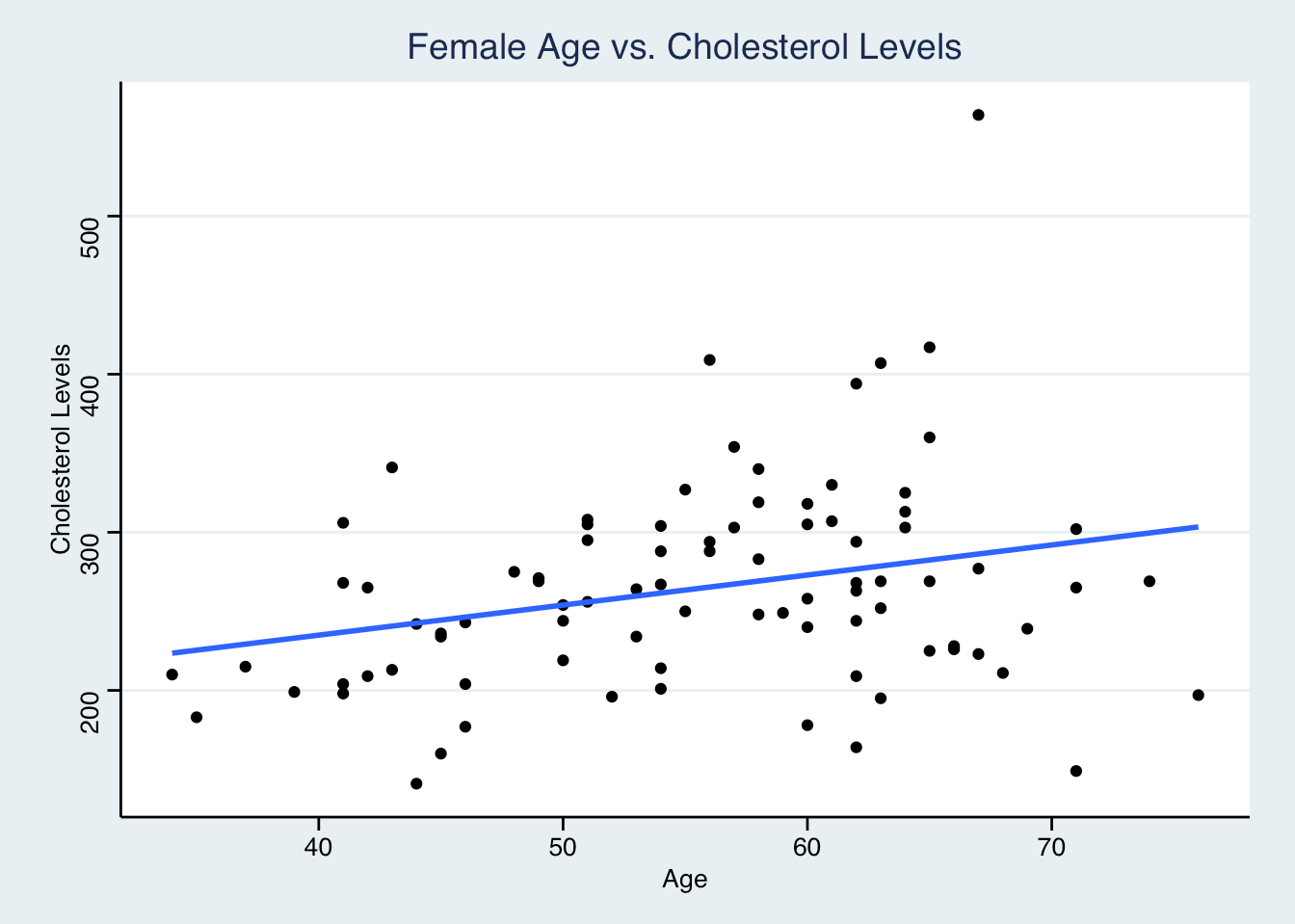

Problem 7. (3 points) Is there a relationship between age and cholesterol levels? Make a plot that shows this relationship separated by gender (hint: use faceting or make two plots). Be sure to add a line of best fit (linear regression line). *Both male and female have a relationship between age and cholesterol levels, the older the person is, the higher the cholesterol levels*

```
heart %>%
  filter(gender=="male") %>%
  ggplot(aes(x=age, y=chol))+
  geom_point()+
  geom_smooth(method = lm, se=F)+
  labs(title = "Male Age vs. Cholesterol Levels",
       x="Age",
       y="Cholesterol Levels",
       alpha=0.6)+
  theme(plot.title = element_text(size=rel(1.5), hjust=0.5))+
  theme_stata()
```

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

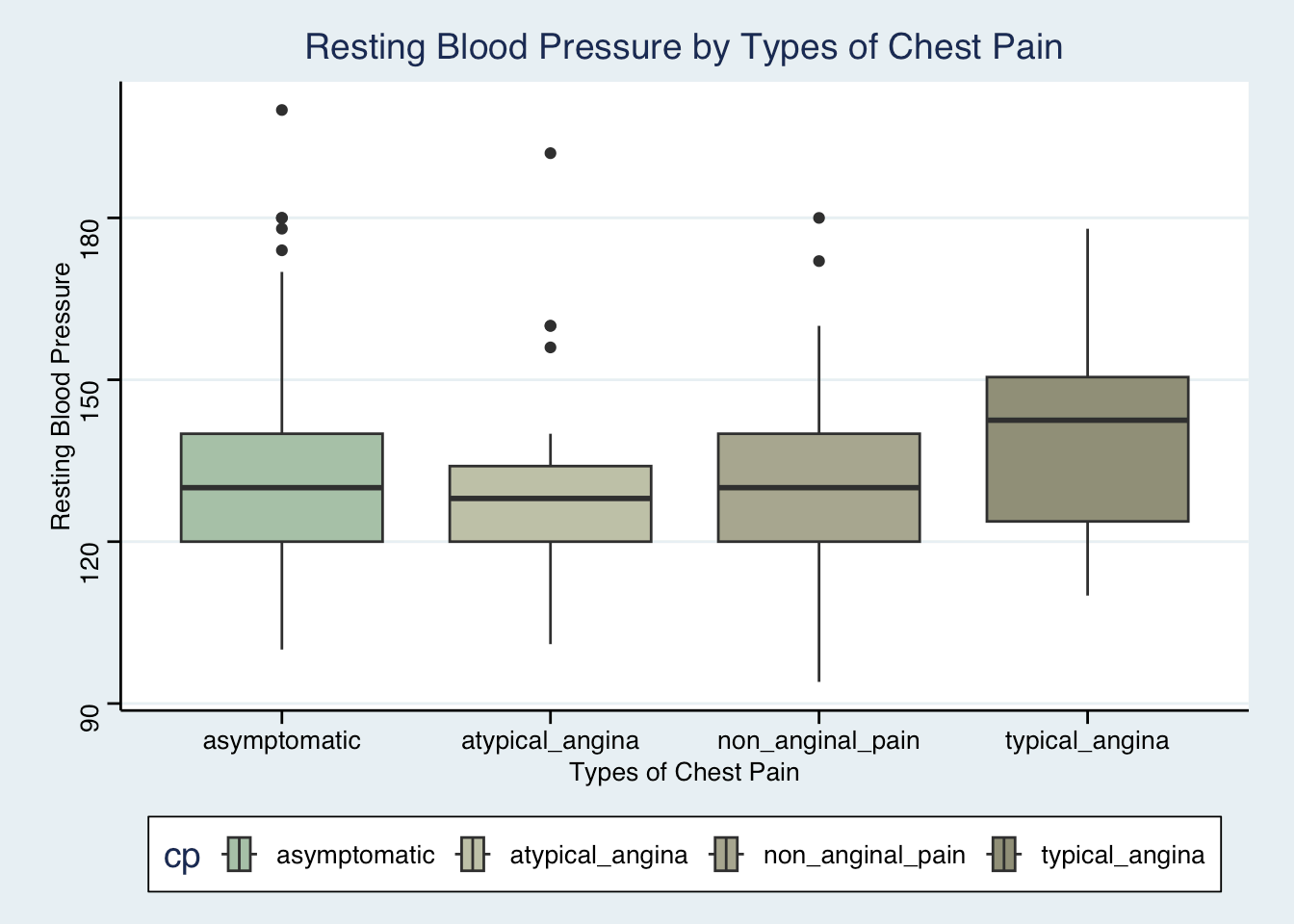
```
heart %>%
  filter(gender=="female") %>%
  ggplot(aes(x=age, y=chol))+
  geom_point()+
  geom_smooth(method = lm, se=F)+
  labs(title = "Female Age vs. Cholesterol Levels",
       x="Age",
       y="Cholesterol Levels",
       alpha=0.6)+
  theme(plot.title = element_text(size=rel(1.5), hjust=0.5))+
  theme_stata()
```

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



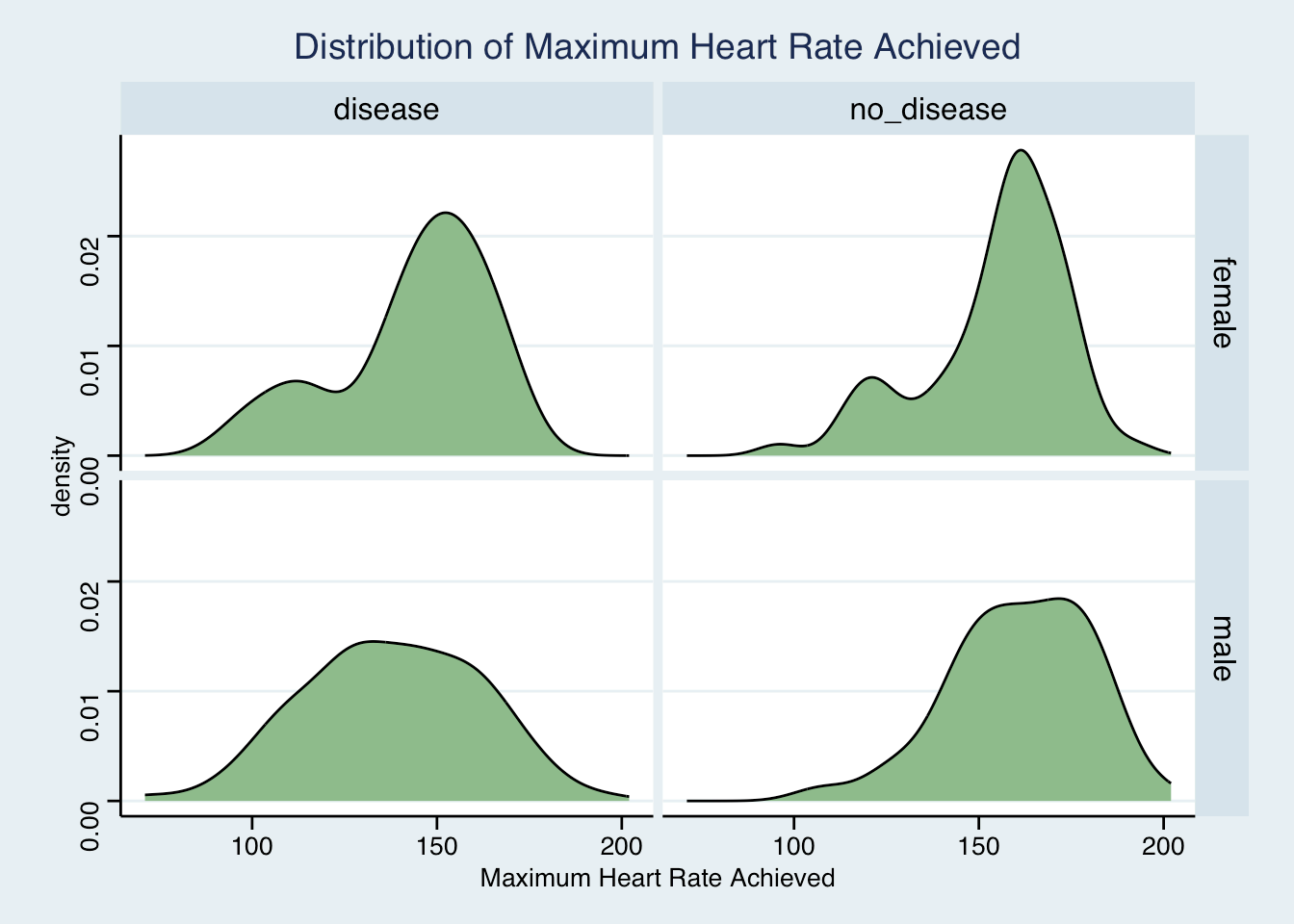

Problem 8. (3 points) What is the range of resting blood pressure for participants by type of chest pain? Make a plot that shows this information. _The range is

```
heart %>%
  ggplot(aes(x=trestbps, y=cp, fill=cp))+
  geom_boxplot()+
  labs(title = "Resting Blood Pressure by Types of Chest Pain",
       x="Resting Blood Pressure",
       y="Types of Chest Pain",
       alpha=0.6)+
  theme(plot.title = element_text(size=rel(1.5), hjust=0.5),
        legend.position = "bottom")+
  theme_stata()+
  scale_fill_manual(values=my_palette)+
  coord_flip()
```




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. *Female has a higher peak on both disease and no disease than that of males; but besides the male_disease graphs, all graphs have a peak above 150.*

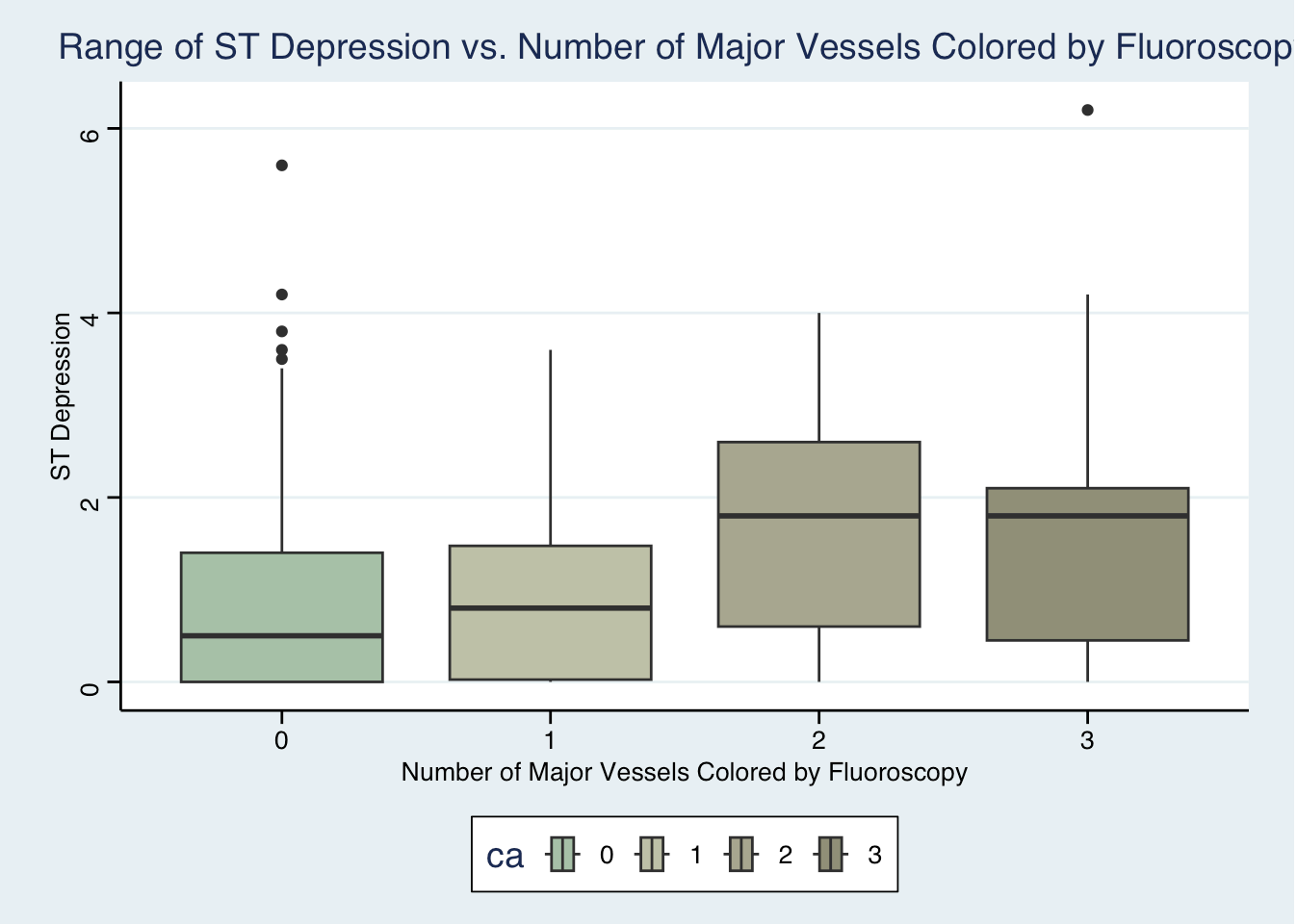
```
heart %>%
  ggplot(aes(x=thalach))+
  geom_density(fill= "darkseagreen")+
  facet_grid(~gender~target)+
  labs(title = "Distribution of Maximum Heart Rate Achieved",
       x="Maximum Heart Rate Achieved",
       alpha=0.6)+
  theme(plot.title = element_text(size=rel(1.5), hjust=0.5))+
  theme_stata()
```




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?)

*When ca is 0, it has a range of 0~1.7; when ca is 1, it has a range of 0~ 1.73; when ca is 2, its range is 0.75~2.5; when ca is 3, its range is 0.6~2.3*

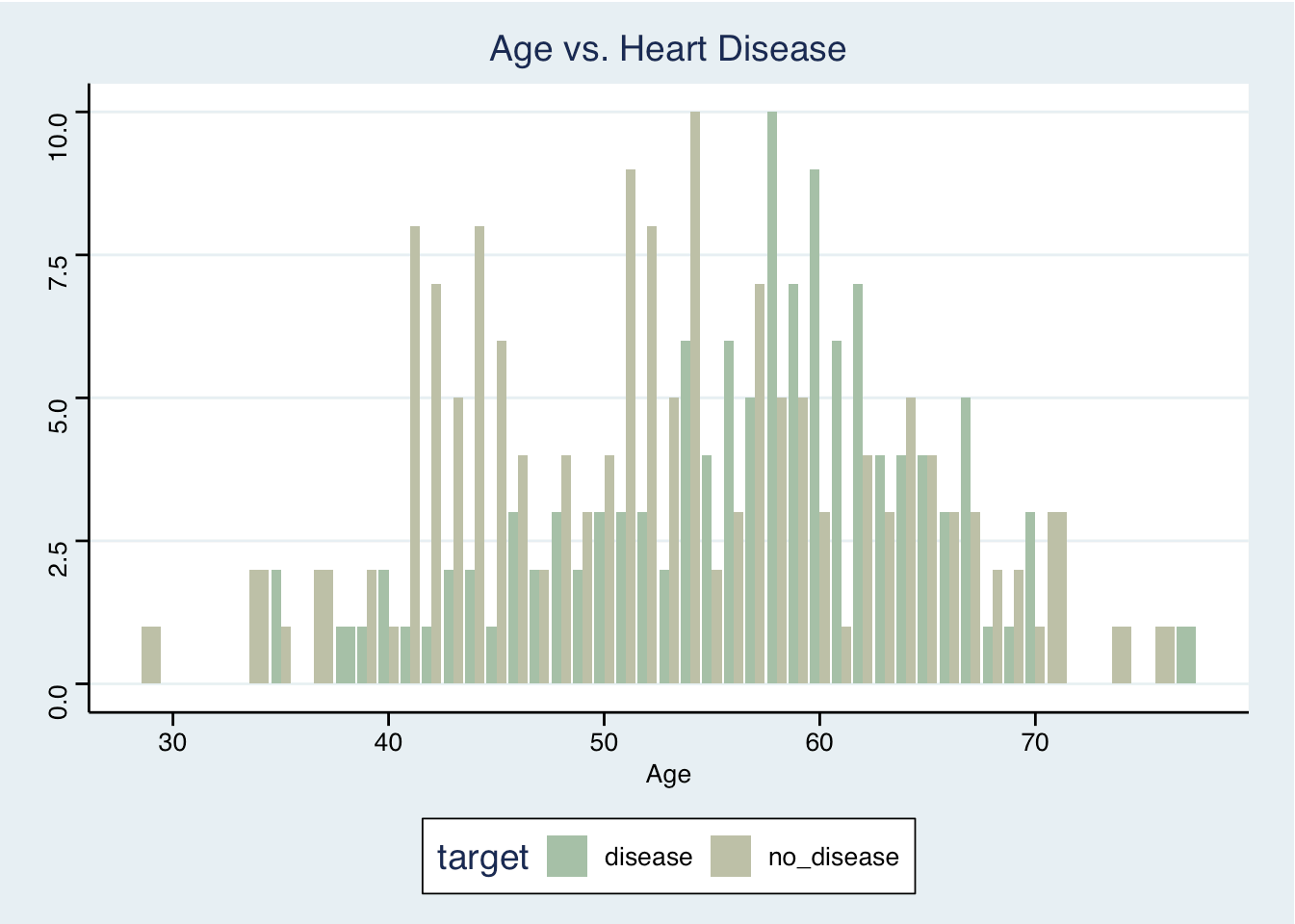
```
heart %>%
  mutate(ca=as.factor(ca)) %>%
  ggplot(aes(x=oldpeak, y=ca, group = ca, fill = ca))+
  geom_boxplot()+
  labs(title = "Range of ST Depression vs. Number of Major Vessels Colored by Fluoroscop
y",
       x="ST Depression",
       y="Number of Major Vessels Colored by Fluoroscopy",
       alpha=0.6)+
  theme(plot.title = element_text(size=rel(1.5), hjust=0.5),
        legend.position = "bottom")+
  theme_stata()+
  scale_fill_manual(values=my_palette)+
  coord_flip()
```




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.

*Yes, at around age 57, I can see that the number of people with heart disease surpass the number of people without heart disease for the first time in the data*

```
heart %>%
  ggplot(aes(x=age, fill=target))+
  geom_bar(position = "dodge")+
  labs(title = "Age vs. Heart Disease",
       x="Age",
       y=NULL)+
  theme_stata()+
  scale_fill_manual(values=my_palette)
```




# 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.