Midterm 2 W25

Luc-Tanton Tran

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 %>%
  group_by(gender) %>%
  summarize(n = n())
```

```
## # 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(avg_age = mean(age))
```

```
## # A tibble: 2 × 2
##   gender avg_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.

```
heart %>%
  group_by(gender, target) %>%
  summarize(n = n()) %>%
  pivot_wider(names_from = "target",
              values_from = "n")
```

```
## `summarise()` has grouped output by 'gender'. You can override using the
## `.groups` argument.
```

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

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

```
heart %>%
  group_by(gender, target) %>%
  summarize(n = n()) %>%
  group_by(gender) %>%
  summarize(n,total = sum(n), percentage = n/total*100, .groups = 'keep')
```

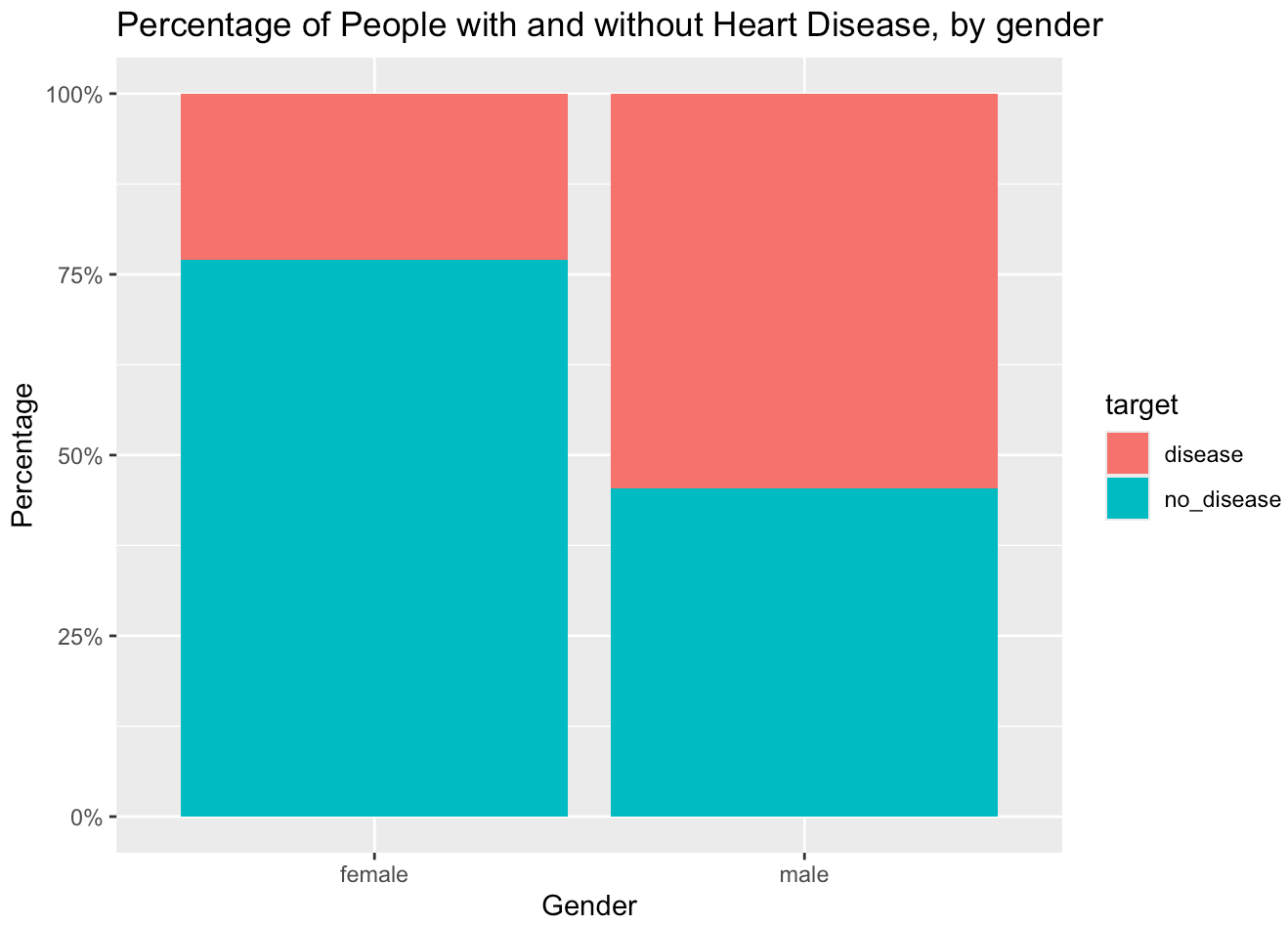
```
## `summarise()` has grouped output by 'gender'. You can override using the
## `.groups` argument.
```

```
## Warning: Returning more (or less) than 1 row per `summarise()` group was deprecated i
n
## dplyr 1.1.0.
## ℹ Please use `reframe()` instead.
## ℹ When switching from `summarise()` to `reframe()`, remember that `reframe()`
##   always returns an ungrouped data frame and adjust accordingly.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## # A tibble: 4 × 4
## # Groups:   gender [2]
##   gender     n total percentage
##   <chr>  <int> <int>      <dbl>
## 1 female    20    87       23.0
## 2 female    67    87       77.0
## 3 male     100   183       54.6
## 4 male      83   183       45.4
```

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)+
  labs(title = "Percentage of People with and without Heart Disease, by gender",
       x = "Gender",
       y = "Percentage")
```



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).

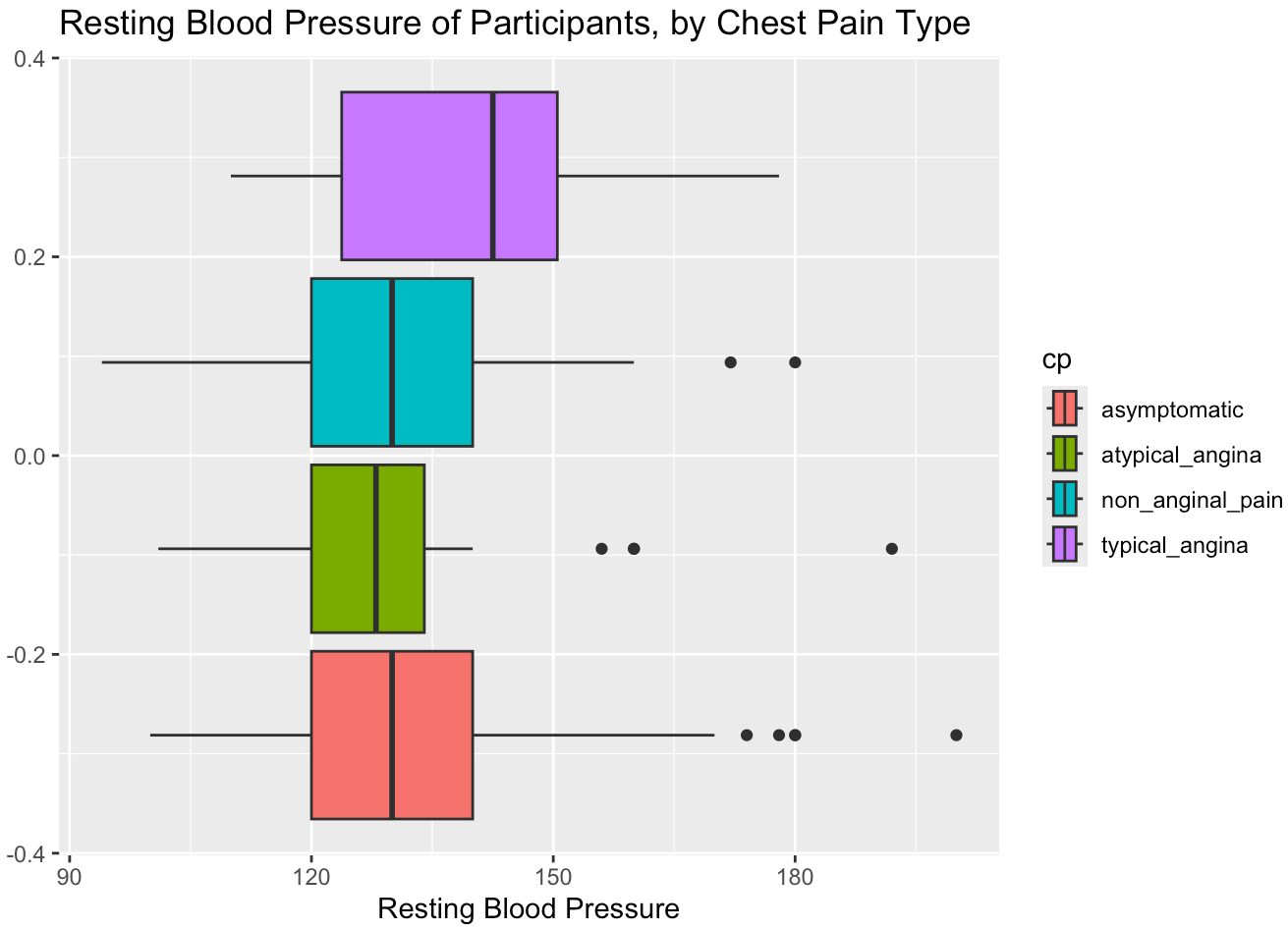
There seems to be a positive relationship between age and cholesterol levels, though it seems slightly more prominent in females than in males.

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

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

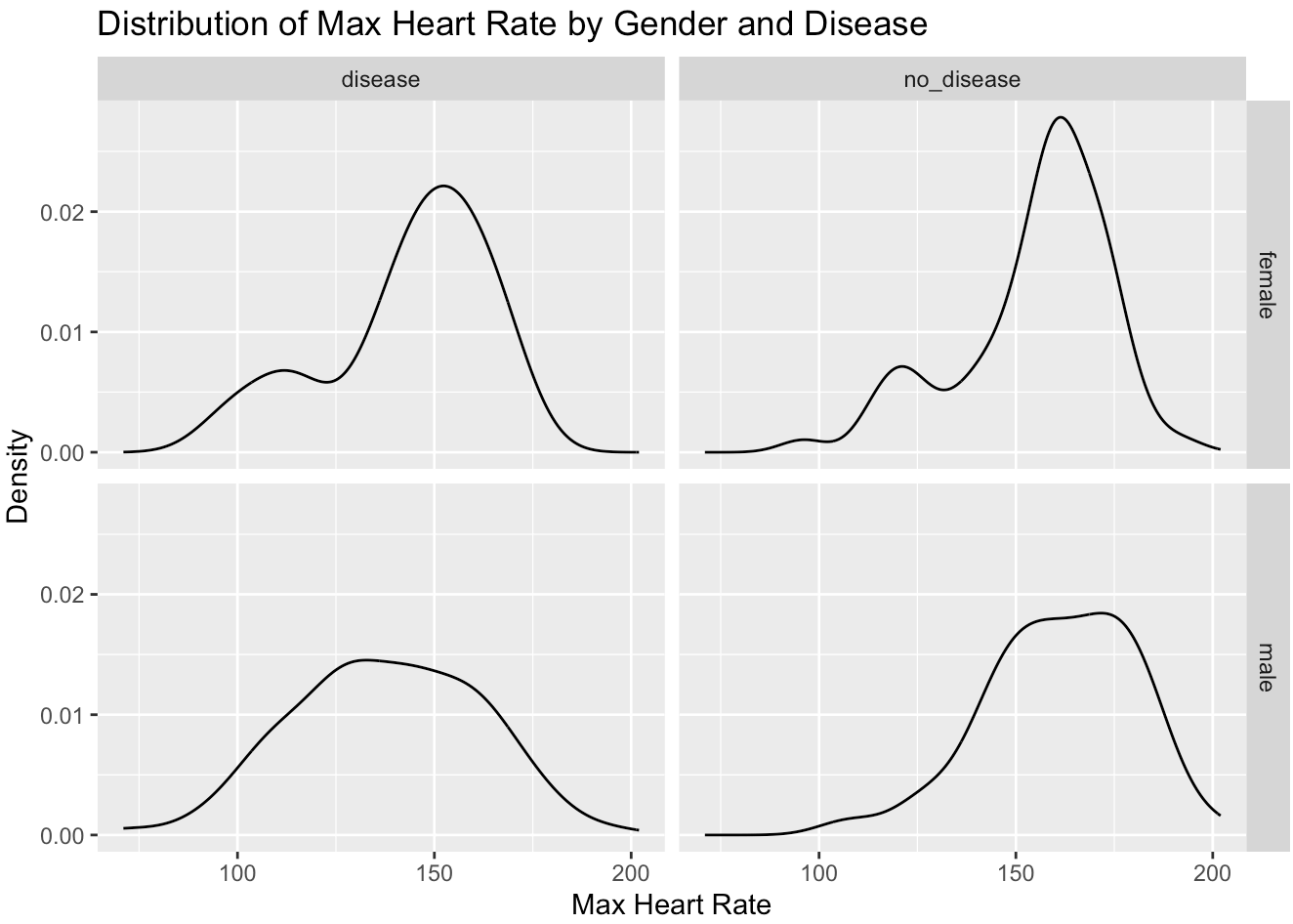
Problem 8. (3 points) What is the range of resting blood pressure for participants by type of chest pain? Make a plot that shows this information.

```
heart %>%
  ggplot(aes(trestbps, fill = cp)) +
  geom_boxplot()+
  labs(
    title = "Resting Blood Pressure of Participants, by Chest Pain Type",
    x = "Resting Blood Pressure",
    y = NULL
  )
```



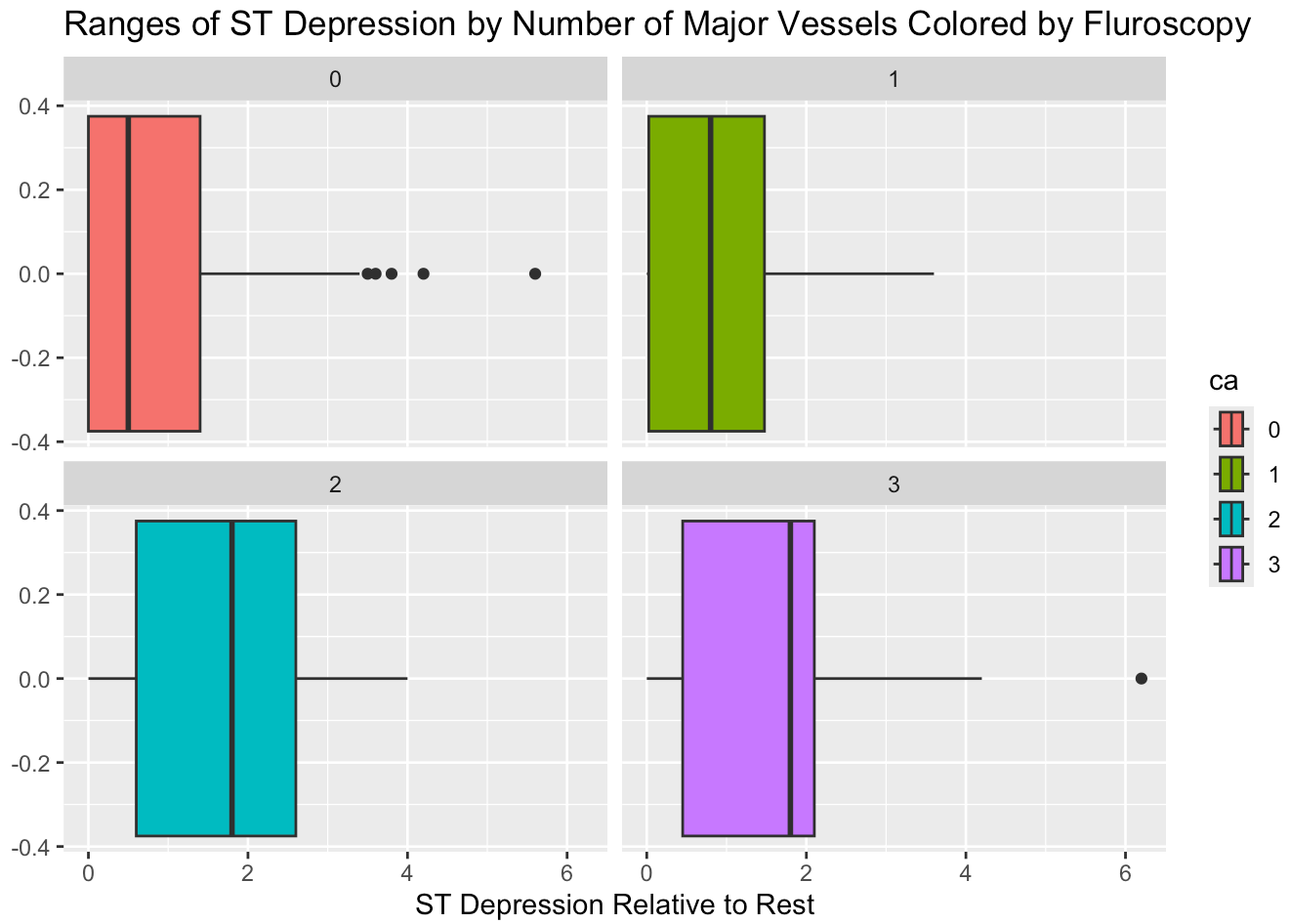
Problem 9. (4 points) What is the distribution of maximum heart rate achieved, separated by gender and whether or not the patient has heart disease? Make a plot that shows this information- you must use faceting.

```
heart %>%
  ggplot(aes(x = thalach))+
  geom_density()+
  facet_grid(gender~target)+
  labs(title = "Distribution of Max Heart Rate by Gender and Disease",
       x = "Max Heart Rate",
       y = "Density")
```



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 %>%
  mutate(ca = as.factor(ca)) %>%
  ggplot(aes(x = oldpeak, group = ca, fill = ca))+
  geom_boxplot()+
  facet_wrap(ca~.)+
  labs(title = "Ranges of ST Depression by Number of Major Vessels Colored by Fluroscop
y",
       x = "ST Depression Relative to Rest")
```

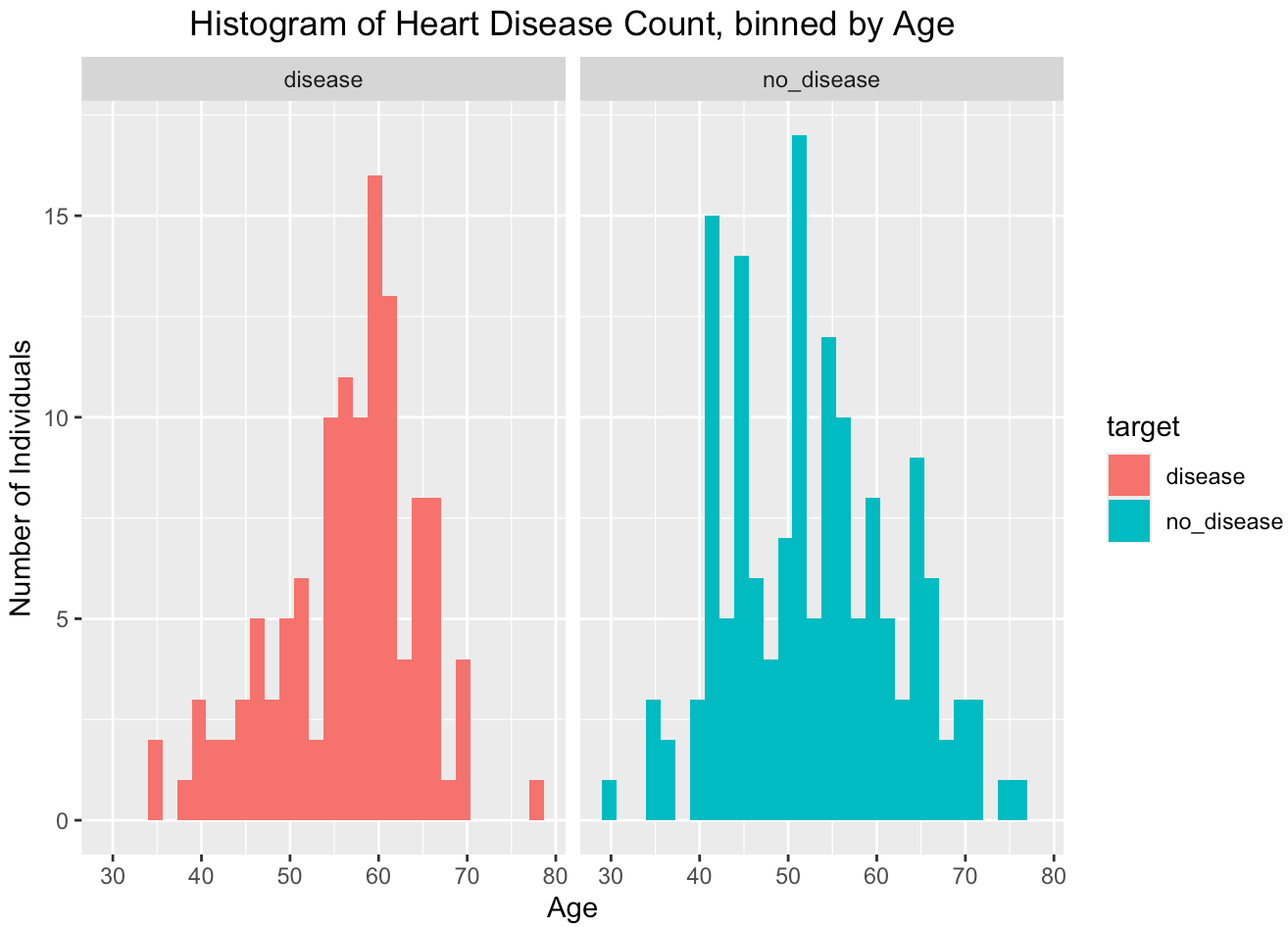


Problem 11. (4 points) Is there an age group where we see increased prevalence of heart disease? Make a plot that shows the number of participants with and without heart disease by age group.

There seems to be a peak around the 55-60 year old age group that has a relatively higher risk than the other age groups, where it suddenly jumps up (from below 5 to about 10), with the highest number around 59 years old (above 15 individuals).

```
heart %>%
  ggplot(aes(x = age, fill = target))+
  geom_histogram()+
  facet_wrap(target~.)+
  labs(title = "Histogram of Heart Disease Count, binned by Age",
       x = "Age",
       y = "Number of Individuals")+
  theme(plot.title = element_text(hjust= 0.5))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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.