

## Intermediate ggplot2


STA 418/518 - Statistical Computing and Graphics with R

Andrew DiLernia

Complete the following activity using R Markdown and then submit your .Rmd file and output Word, PDF, or HTML document on Blackboard.

### Learning Objectives

1. Exploratory data analysis
2. Use of `geom_point()` and `geom_smooth()`
3. Complete themes in `ggplot2`
4. Customizing theme elements in `ggplot2`
5. Creating and customizing bar charts in `ggplot2`
6. Faceting and combining plots with the `patchwork` package

➡ First, create a new R Markdown document with *File > New File > R Markdown...* Knit it by clicking the  Knit button (top left).

- Knit it by using the appropriate keyboard shortcut (**Mac:** *Command + Shift + K*, **Windows:** *Ctrl + Shift + K*).
- Load packages necessary for this activity using the code below. Note: you may need to install packages beforehand by using the `install.packages()` function and the package name in quotes.

```
library(riskCommunicator)
library(tidyverse)
library(skimr)
library(knitr)
library(ggthemes)
library(patchwork)
```

### The Framingham Heart Study

Launched in 1948, the [Framingham Heart Study](#) (FHS) is a longitudinal, multigenerational study consisting of participants from Framingham, Massachusetts. The main purpose of the study was to examine the risks and causes of cardiovascular disease. Over time, the FHS has evolved to analyze familial trends of cardiovascular disease and other diseases, while

simultaneously collecting genetics data on multiple generations descended from the original study participants. To improve generalizability and representativeness of the study, the FHS has also expanded to include a more diverse set of participants to improve the understanding of the etiology of cardiovascular disease and other diseases for the broader population.



## Framingham Heart Study

Three Generations of Dedication

Image from <https://www.framinghamheartstudy.org/>

→ First, let's load the FHS data set from the `riskCommunicator` package and explore some high-level characteristics of the data set using the code below:

```
data(framingham, package = "riskCommunicator")
```

```
?framingham
```

```
glimpse(framingham)
```

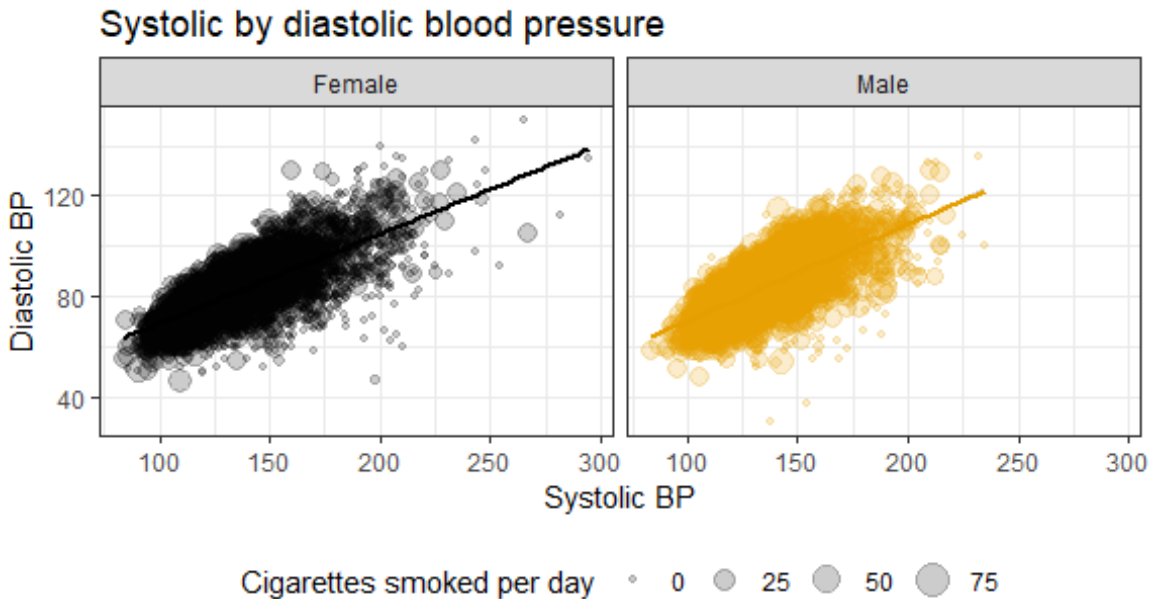
→ Select the first 10 variables from the Framingham dataset and store it as a new data frame called `framinghamSub` using the `select()` function. Also, update the `SEX` variable to have the values "Male" and "Female" rather than 1 and 2, and the `CURSMOKE` variable to have the values "Yes" and "No" rather than 1 and 0 using the `mutate()` and `case_when()` functions. This should be your new dataset to be used for the rest of the assignment.

→ Use the `skim()` function from the `skimr` package to explore other characteristics of the subset of the data.

→ Make a scatter plot between diastolic (`DIABP`) and systolic (`SYSBP`) blood pressure with a "facet" by the sex of the participant (`SEX`). Also manually set the `alpha` aesthetic to be 0.2. After the next few bullets is an example.

→ Also include the size of the data points as mapped by the number of cigarettes smoked per day (`CIGPDAY`), add a color-blind friendly palette for coloring the points based on the sex of each participant, and position the legend at the bottom of the plot.

→ Add a line of best fit corresponding to a simple linear regression model fit separately for males and females using `geom_smooth()`.

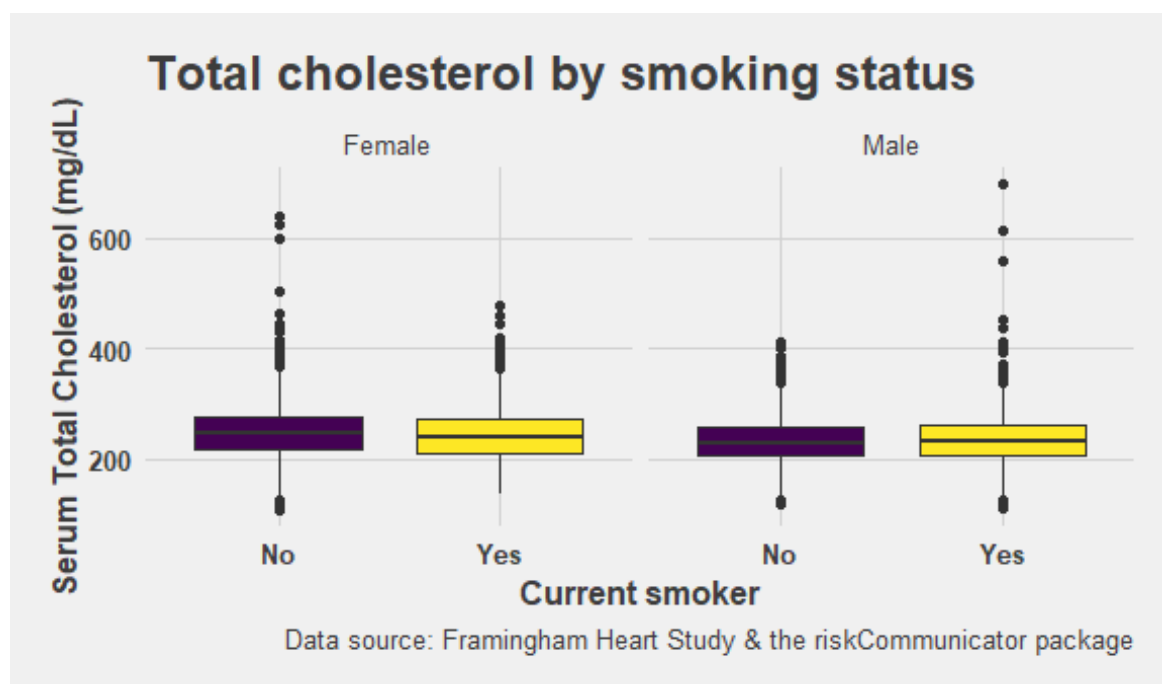


Data source: Framingham Heart Study & the riskCommunicator package

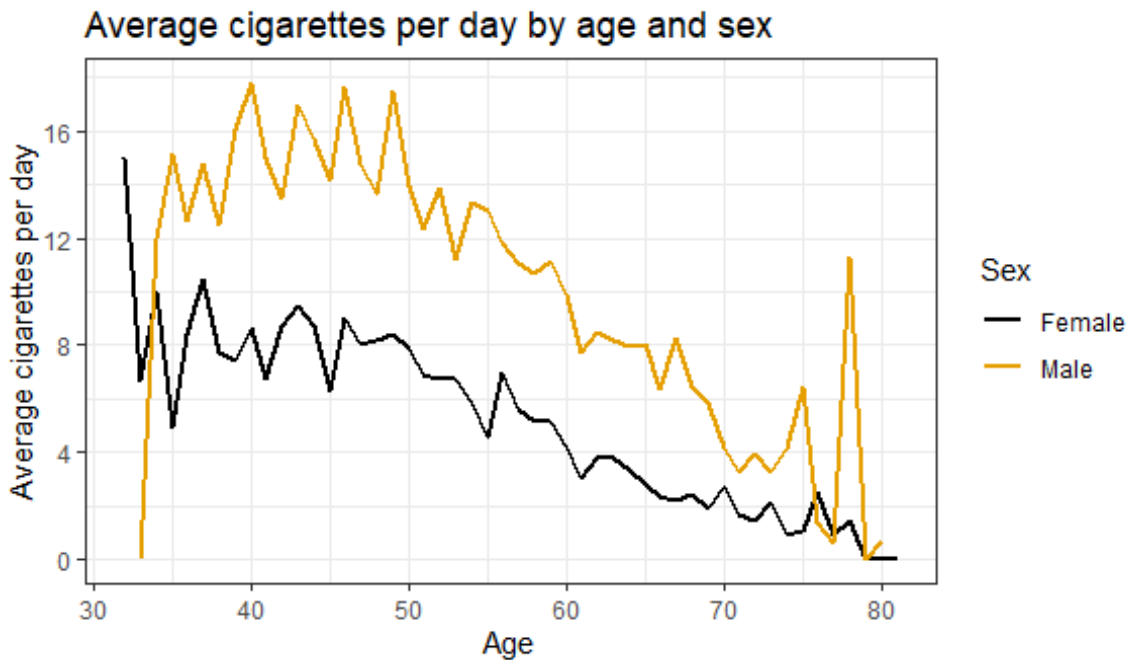
➡ Next, create a side-by-side box-plot where the y-axis is total cholesterol (TOTCHOL) and the x-axis is current smoking status (CURSMOKE). Make all axis and title text bold in the plot.

➡ Add a complete theme from `ggthemes`, color the boxes based on smoking status, remove the legend, and make the axis titles bold and change the font size as well.

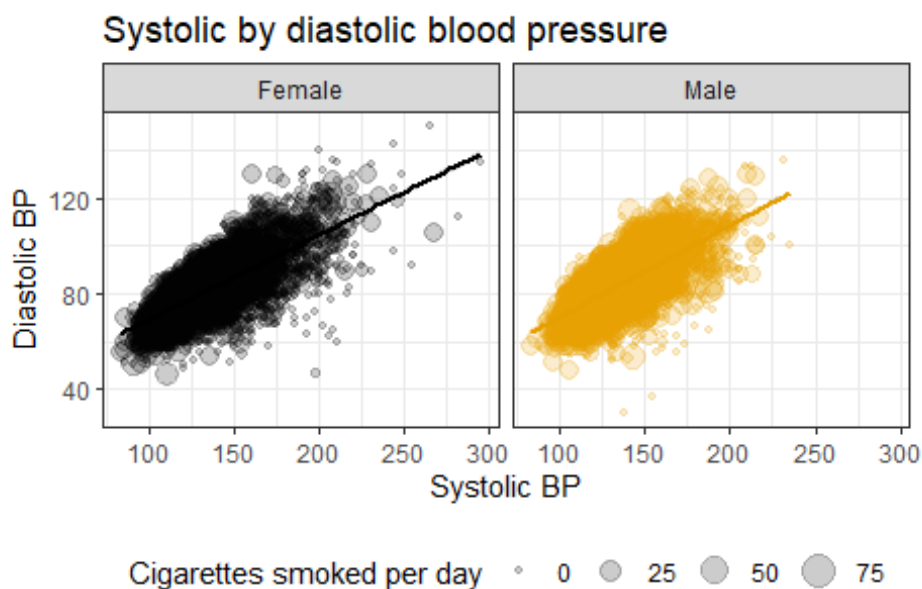
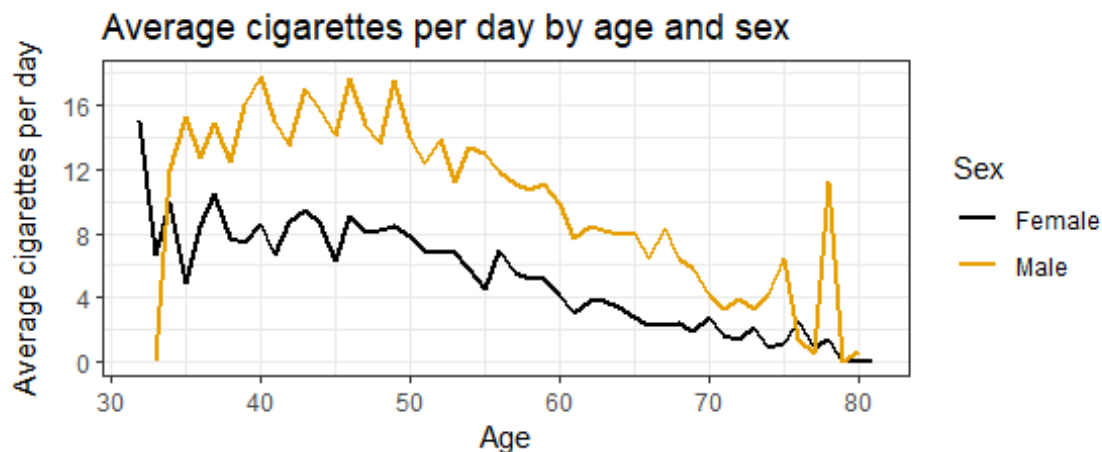
➡ In a new plot, modify the side-by-side box-plots we created to be faceted by the sex of the participant using the `facet_grid()` function and columns to break up the subplots.



- Make a line graph that shows the average cigarettes per day (CIGPDAY) by age (AGE), with separate lines by the sex of the participant (SEX).
- Apply a complete theme to the plot, and have the axis show the breaks at 0, 4, 8, 12, and 16 cigarettes per day.



- Combine the line chart and the faceted scatter plots together into a single graphic using the patchwork package, with 1 plot per row and the line chart on top.



Data source: Framingham Heart Study & the riskCommunicator package

According to [Johns Hopkins Medicine](#), one way to bin / categorize total cholesterol levels is as Normal (<200 mg/dL), Borderline high (200 to 239 mg/dL), or High (> 240 mg/dL). Using the code below, we can create a new variable, `CholesterolCat`, indicating whether each participant's total cholesterol is considered, Normal, Borderline high, or High.

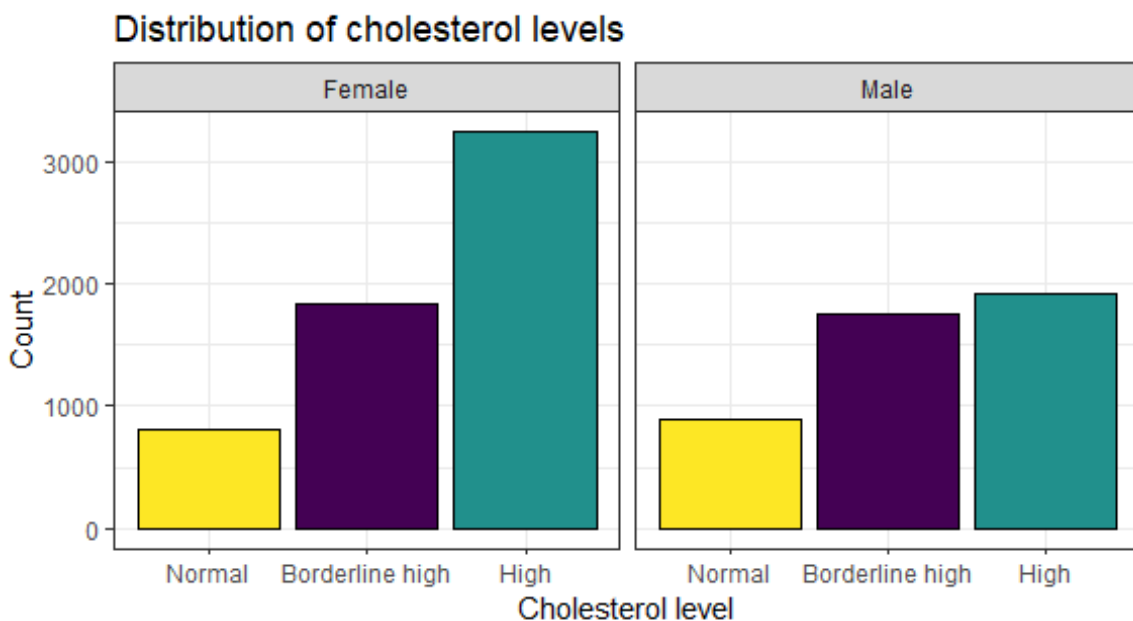
```
framinghamSub <- framinghamSub %>%
  mutate(CholesterolCat = case_when(TOTCHOL < 200 ~ "Normal",
                                     TOTCHOL >= 200 & TOTCHOL < 240 ~
"Borderline high",
                                     TOTCHOL > 240 ~ "High",
                                     TRUE ~ as.character(NA)))
```

➡ Create a bar chart displaying the number of participants falling in each cholesterol category based on Johns Hopkins' definitions using `geom_bar()`. Also, remove people under

40 and those without recorded cholesterol levels (missing values for CholesterolCat) from the plot by using the code `filter(AGE >= 40, !is.na(CholesterolCat))` when piping the data into each subsequent `ggplot()` call.

→ Recreate the bar chart, this time reordering the categories to show Normal, Borderline high, and then High from left to right using the `fct_relevel()` function.

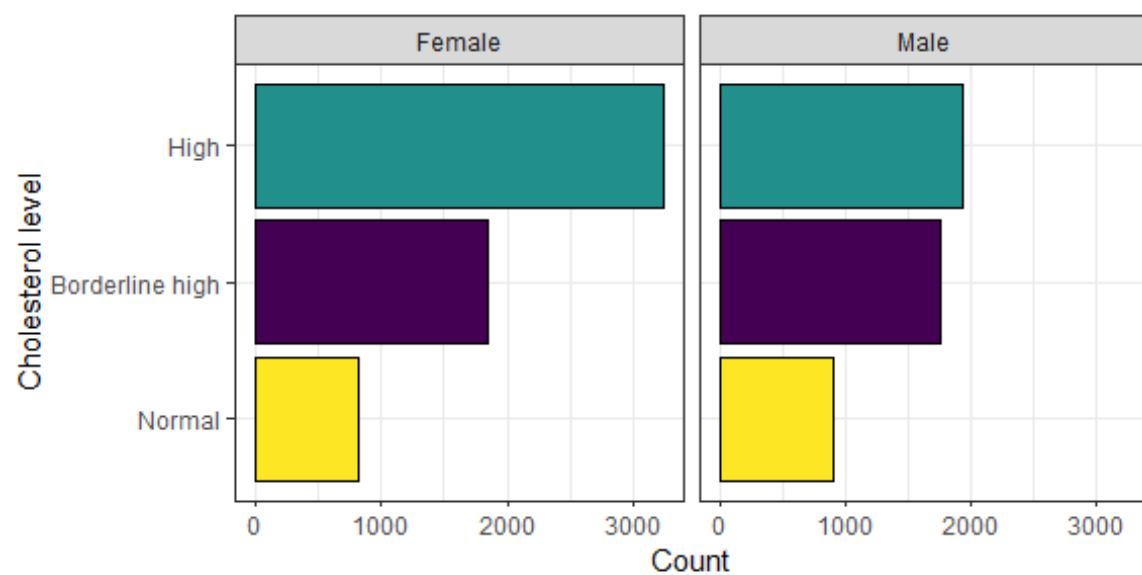
→ Change the color of the inside of the bars based on the cholesterol category using a color-blind friendly palette, make the outline of the bars black in color, facet by the sex of the participant, and remove the legend.



Data source: Framingham Heart Study & the riskCommunicator package

→ Lastly, use the `coord_flip()` function to turn the bar chart into a horizontal bar chart instead.

## Distribution of cholesterol levels



Data source: Framingham Heart Study & the riskCommunicator package