Intermediate ggplot2

Aditya Dube

Wednesday, May 17, 2023

Loading Packages

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

First, let's load the FHS data set from the riskCommunicator package

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

```
## Rows: 11,627
## Columns: 39
## $ RANDID   <dbl> 2448, 2448, 6238, 6238, 6238, 9428, 9428, 10552, 10552, 11252~
## $ SEX      <dbl> 1, 1, 2, 2, 2, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1~
## $ TOTCHOL  <dbl> 195, 209, 250, 260, 237, 245, 283, 225, 232, 285, 343, NA, 22~
## $ AGE      <dbl> 39, 52, 46, 52, 58, 48, 54, 61, 67, 46, 51, 58, 43, 49, 55, 6~
## $ SYSBP    <dbl> 106.0, 121.0, 121.0, 105.0, 108.0, 127.5, 141.0, 150.0, 183.0~
## $ DIABP    <dbl> 70.0, 66.0, 81.0, 69.5, 66.0, 80.0, 89.0, 95.0, 109.0, 84.0, ~
## $ CURSMOKE <dbl> 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 0~
## $ CIGPDAY  <dbl> 0, 0, 0, 0, 0, 20, 30, 30, 20, 23, 30, 30, 0, 0, 0, 0, 0, 20,~
## $ BMI      <dbl> 26.97, NA, 28.73, 29.43, 28.50, 25.34, 25.34, 28.58, 30.18, 2~
## $ DIABETES <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0~
## $ BPMEDS   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0~
## $ HEARTRTE <dbl> 80, 69, 95, 80, 80, 75, 75, 65, 60, 85, 90, 74, 77, 120, 86, ~
## $ GLUCOSE  <dbl> 77, 92, 76, 86, 71, 70, 87, 103, 89, 85, 72, NA, 99, 86, 81, ~
## $ educ     <dbl> 4, 4, 2, 2, 2, 1, 1, 3, 3, 3, 3, 3, 2, 2, 2, 1, 1, 2, 2, 2, 1~
## $ PREVCHD  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0~
## $ PREVAP   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0~
## $ PREVMI   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ PREVSTRK <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ PREVHYP  <dbl> 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1, 1, 1, 0, 1, 0, 0, 0, 1~
## $ TIME     <dbl> 0, 4628, 0, 2156, 4344, 0, 2199, 0, 1977, 0, 2072, 4285, 0, 2~
## $ PERIOD   <dbl> 1, 3, 1, 2, 3, 1, 2, 1, 2, 1, 2, 3, 1, 2, 3, 1, 2, 1, 2, 3, 1~
## $ HDLC     <dbl> NA, 31, NA, NA, 54, NA, NA, NA, NA, NA, NA, NA, NA, NA, 46, N~
## $ LDLC     <dbl> NA, 178, NA, NA, 141, NA, NA, NA, NA, NA, NA, NA, NA, NA, 135~
## $ DEATH    <dbl> 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ ANGINA   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0~
## $ HOSPMI   <dbl> 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ MI_FCHD  <dbl> 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0~
## $ ANYCHD   <dbl> 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0~
## $ STROKE   <dbl> 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ CVD      <dbl> 1, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0~
## $ HYPERTEN <dbl> 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ TIMEAP   <dbl> 8766, 8766, 8766, 8766, 8766, 8766, 8766, 2956, 2956, 8766, 8~
## $ TIMEMI   <dbl> 6438, 6438, 8766, 8766, 8766, 8766, 8766, 2956, 2956, 8766, 8~
## $ TIMEMIFC <dbl> 6438, 6438, 8766, 8766, 8766, 8766, 8766, 2956, 2956, 8766, 8~
## $ TIMECHD  <dbl> 6438, 6438, 8766, 8766, 8766, 8766, 8766, 2956, 2956, 8766, 8~
## $ TIMESTRK <dbl> 8766, 8766, 8766, 8766, 8766, 8766, 8766, 2089, 2089, 8766, 8~
## $ TIMECVD  <dbl> 6438, 6438, 8766, 8766, 8766, 8766, 8766, 2089, 2089, 8766, 8~
## $ TIMEDTH  <dbl> 8766, 8766, 8766, 8766, 8766, 8766, 8766, 2956, 2956, 8766, 8~
## $ TIMEHYP  <dbl> 8766, 8766, 8766, 8766, 8766, 8766, 8766, 0, 0, 4285, 4285, 4~
```

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.

```r
framinghamSub <- framingham %>% select(1:10) %>%  mutate(
  SEX = case_when(SEX == 1 ~ "Male",
                  SEX == 2 ~ "Female",
                  TRUE ~ as.character(SEX)),
  CURSMOKE = case_when(
    CURSMOKE == 1 ~ "Yes",
    CURSMOKE == 0 ~ "No",
    TRUE ~ as.character(CURSMOKE)
  )
)
```

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

```r
skim(framinghamSub)
```

Table 1: Data summary

Name	framinghamSub
Number of rows	11627

Number of columns	10
Column type frequency:	
character	2
numeric	8
Group variables	None

Variable type: character

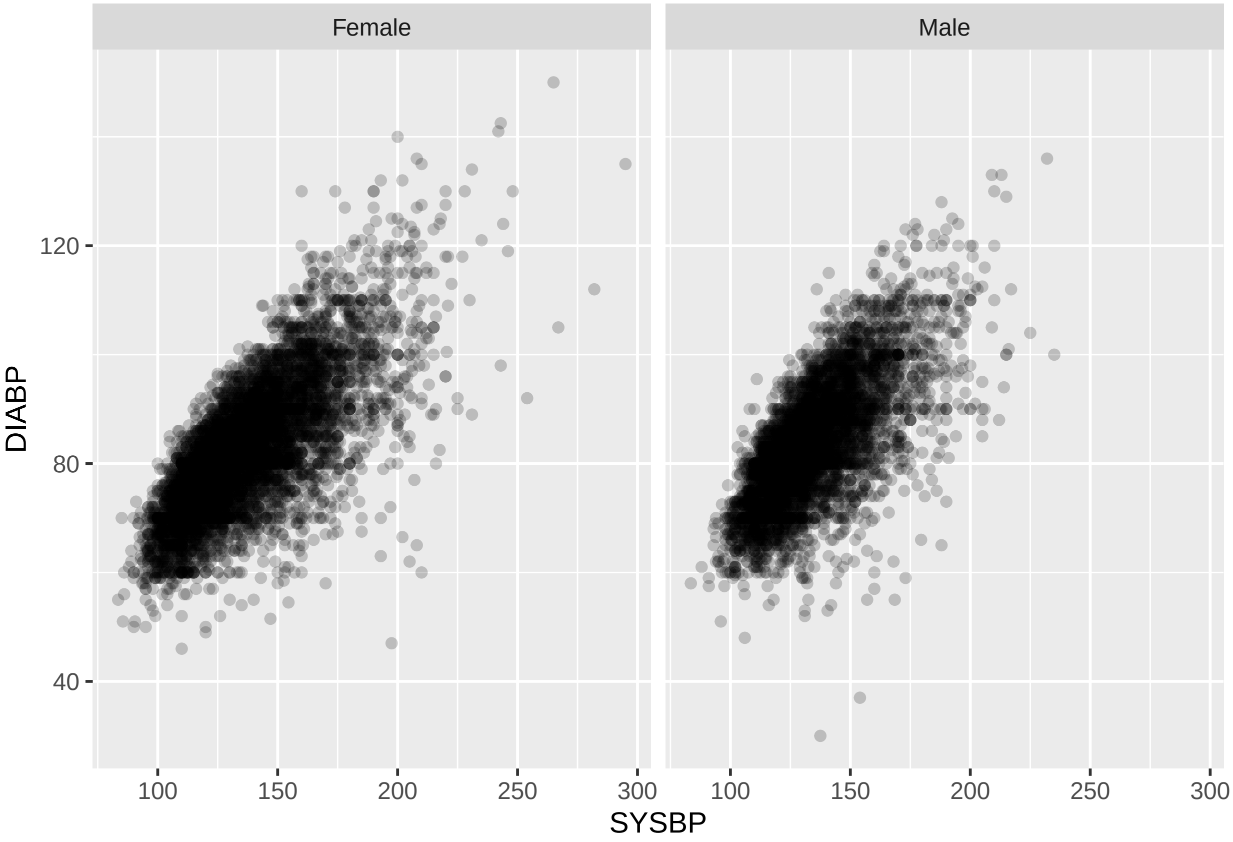
skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
SEX	0	1	4	6	0	2	0
CURSMOKE	0	1	2	3	0	2	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
RANDID	0	1.00	5004740.92	2900877.44	2448.00	2474378.05	5006008.00	7472730.00	9999312.0	
TOTCHOL	409	0.96	241.16	45.37	107.00	210.00	238.00	268.00	696.0	
AGE	0	1.00	54.79	9.56	32.00	48.00	54.00	62.00	81.0	
SYSBP	0	1.00	136.32	22.80	83.50	120.00	132.00	149.00	295.0	
DIABP	0	1.00	83.04	11.66	30.00	75.00	82.00	90.00	150.0	
CIGPDAY	79	0.99	8.25	12.19	0.00	0.00	0.00	20.00	90.0	
BMI	52	1.00	25.88	4.10	14.43	23.09	25.48	28.07	56.8	
DIABETES	0	1.00	0.05	0.21	0.00	0.00	0.00	0.00	1.0	

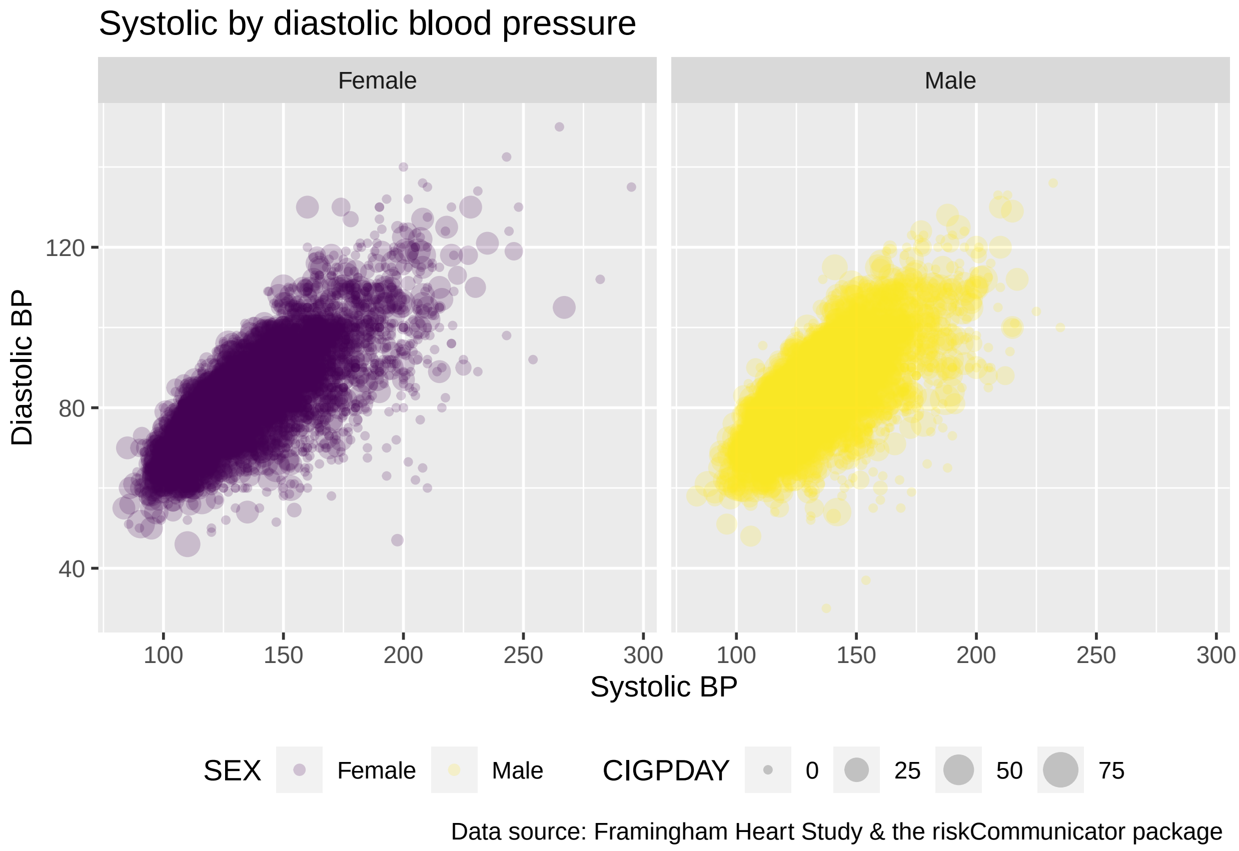
Make a scatter plot between diastolic (DIABP) and systolic (SYSBP) blood pressure with a "facet" by the sex of the participant (SEX). Also manually set the alpha aesthetic to be 0.2. After the next few bullets is an example

```r
ggplot(data = framinghamSub, mapping = aes(x = SYSBP, y = DIABP)) + geom_point(alpha =
 ↪  0.20) + facet_grid(. ~ SEX)
```



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.

```
ggplot(data = framinghamSub,
       mapping = aes(
         x = SYSBP, y = DIABP,
         size = CIGPDAY,
         color = SEX
       )) + geom_point(alpha = 0.20) + facet_grid(. ~ SEX) + scale_color_viridis_d() +
 ↪  labs(
         title = "Systolic by diastolic blood pressure" ,
         x = "Systolic BP" ,
         y = "Diastolic BP",
         caption = "Data source: Framingham Heart Study & the riskCommunicator package "
       ) + theme(legend.position = "bottom")
```

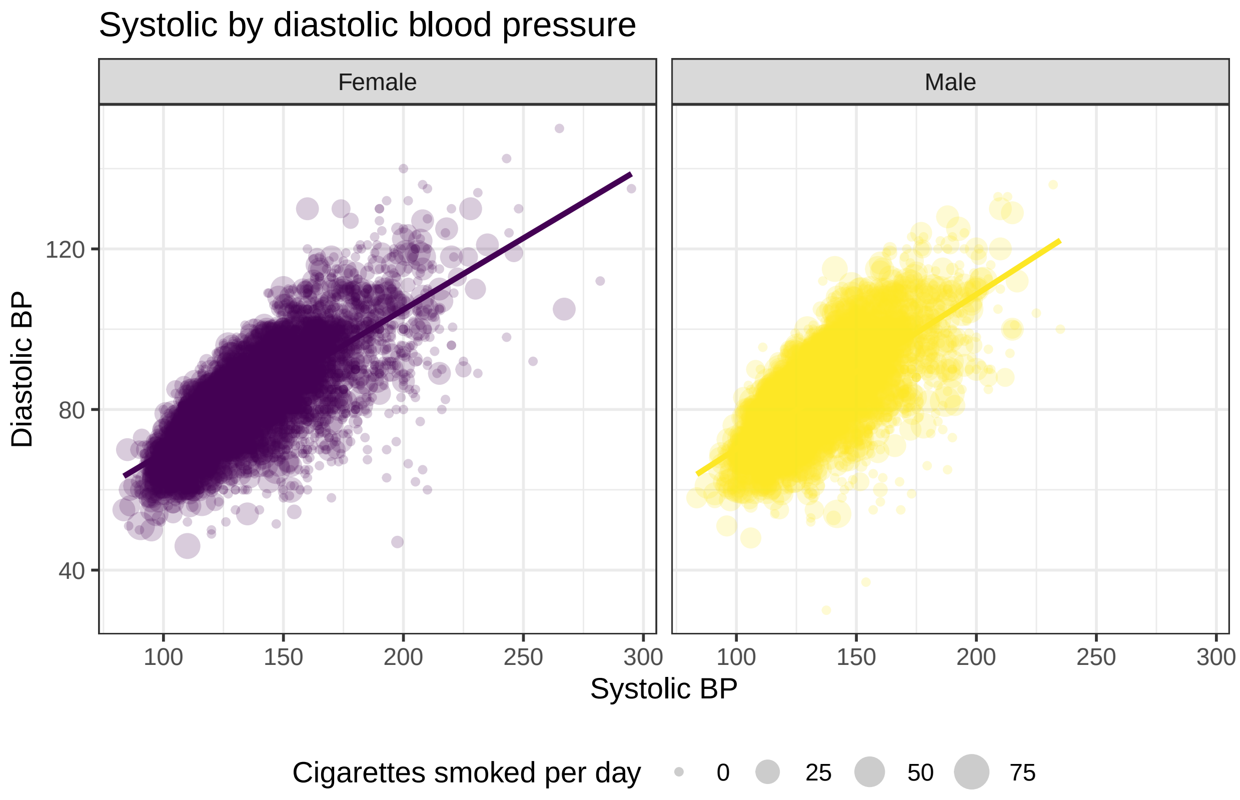

Systolic by diastolic blood pressure

Data source: Framingham Heart Study & the riskCommunicator package

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

```
scatter <- ggplot(data = framinghamSub,
       mapping = aes(
         x = SYSBP,
         y = DIABP,
         size = CIGPDAY,
         color = SEX
       )) + geom_point(alpha = 0.20) + facet_grid(. ~ SEX) + scale_color_viridis_d() +
 ↪  labs(
         title = "Systolic by diastolic blood pressure" ,
         x = "Systolic BP" ,
         y = "Diastolic BP",
         caption = "Data source: Framingham Heart Study & the riskCommunicator package ",
          ↪  size = "Cigarettes smoked per day"
       ) + guides(color = FALSE) + geom_smooth(se = FALSE, method = "lm", size = 1) +
 ↪  theme_bw() + theme(legend.position = "bottom")

scatter
```

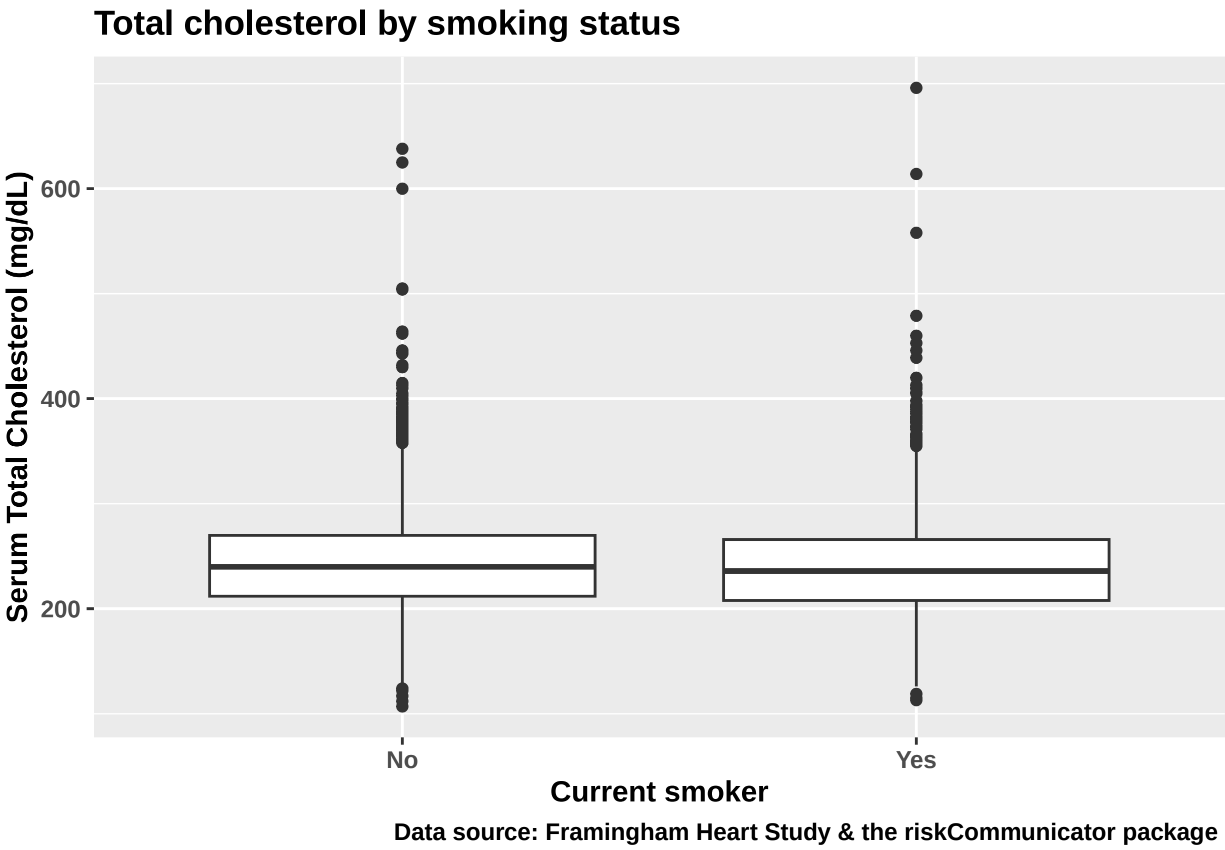


Systolic by diastolic blood pressure

Data source: Framingham Heart Study & the riskCommunicator package

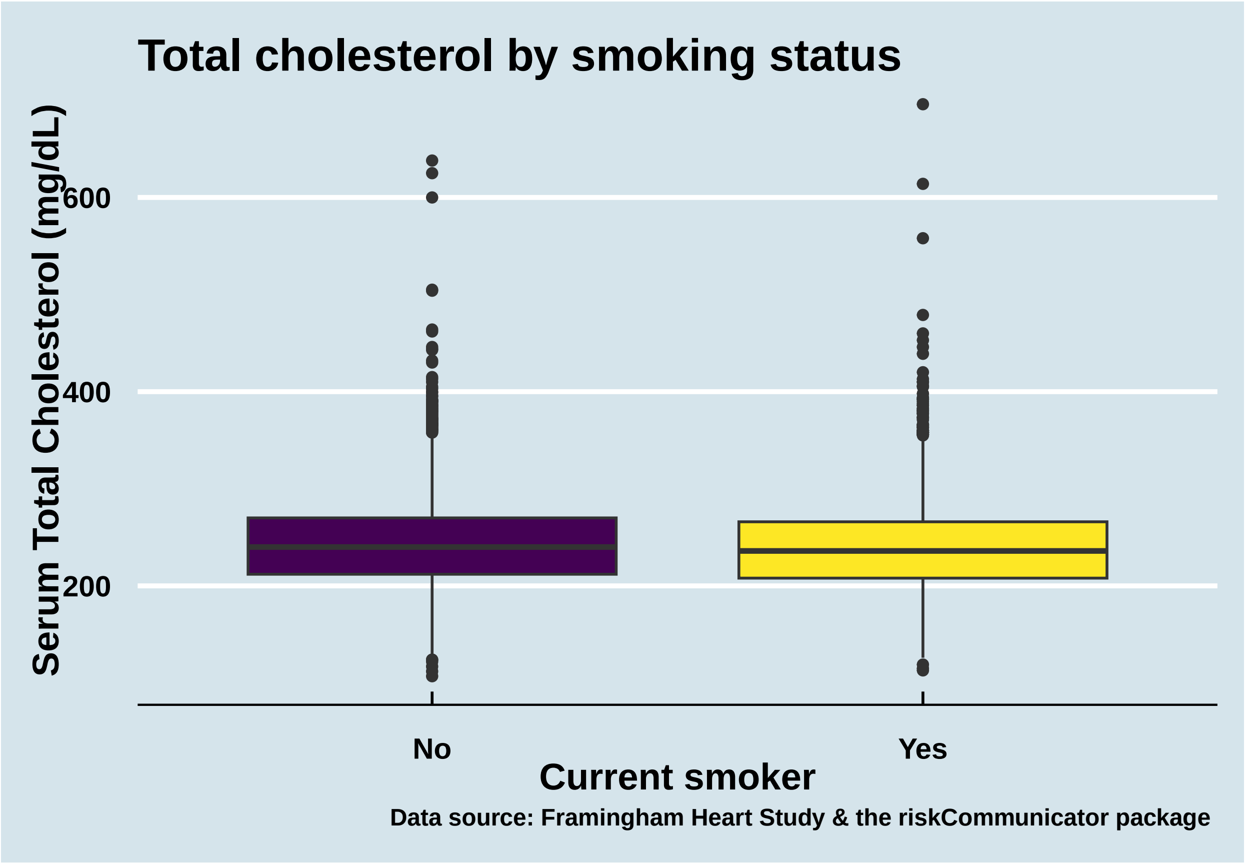
Next, create a side-by-side box-plot where the y-axis is total cholesterol (TOTCHOL) and the x-axis is current smoking status (CURSMOKE). Make all axis and title text bold in the plot.

```
ggplot(
  data = framinghamSub,
  mapping = aes(
    x = CURSMOKE,
    y = TOTCHOL,
  )
) + geom_boxplot() + labs(
  title = "Total cholesterol by smoking status" ,
  x = "Current smoker" ,
  y = "Serum Total Cholesterol (mg/dL)",
  caption = "Data source: Framingham Heart Study & the riskCommunicator package "
)  +
  theme(
    text = element_text(face = "bold"))
```



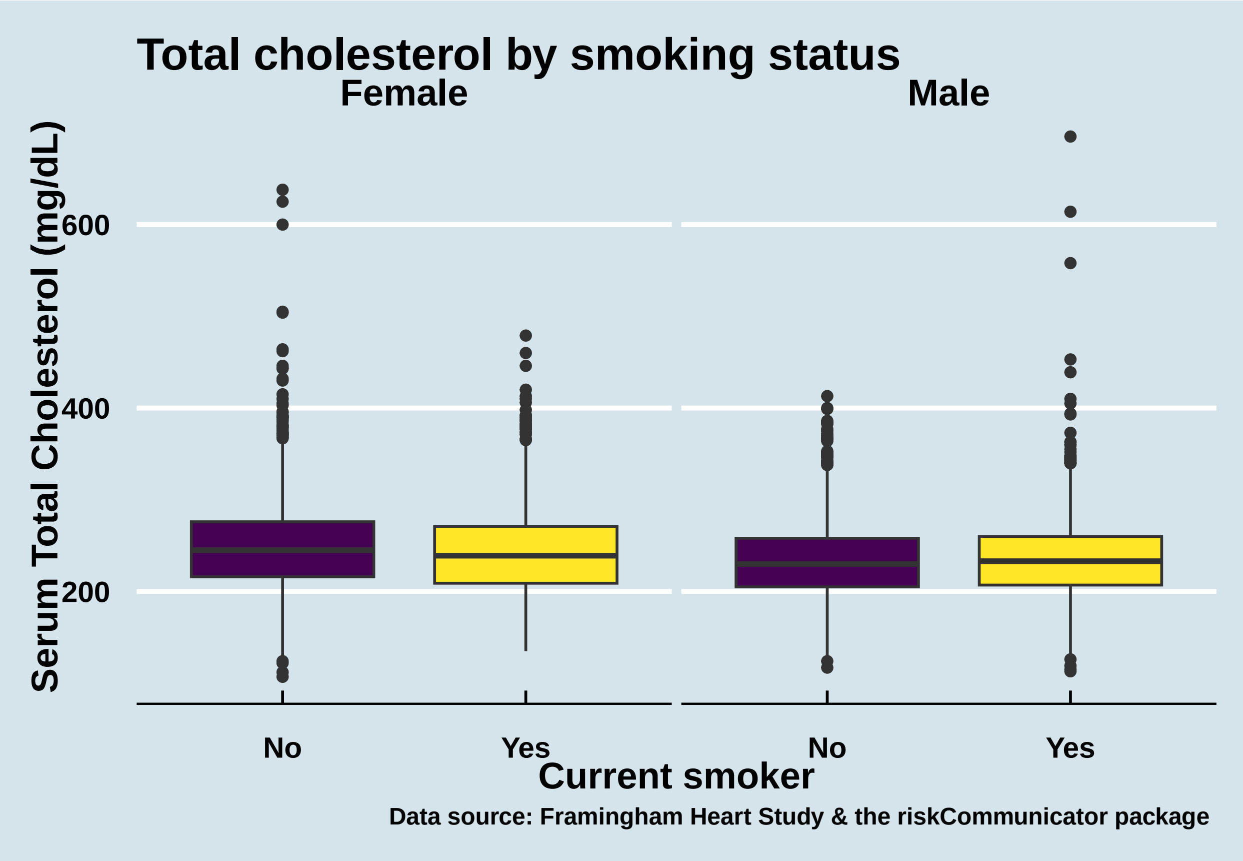
Add a complete theme from ggthemes, color the boxes based on smoking status, remove the legend, and make the axis titles bold and change the font size as well.

```
ggplot(
  data = framinghamSub,
  mapping = aes(
    x = CURSMOKE,
    y = TOTCHOL,
    fill = CURSMOKE
  )
) + geom_boxplot() + scale_fill_viridis_d() + labs(
  title = "Total cholesterol by smoking status" ,
  x = "Current smoker" ,
  y = "Serum Total Cholesterol (mg/dL)",
  caption = "Data source: Framingham Heart Study & the riskCommunicator package "
)   + theme_economist() +
  theme(
    text = element_text(face = "bold"), legend.position = "none",  axis.title =
     ↪ element_text(size = 14))
```



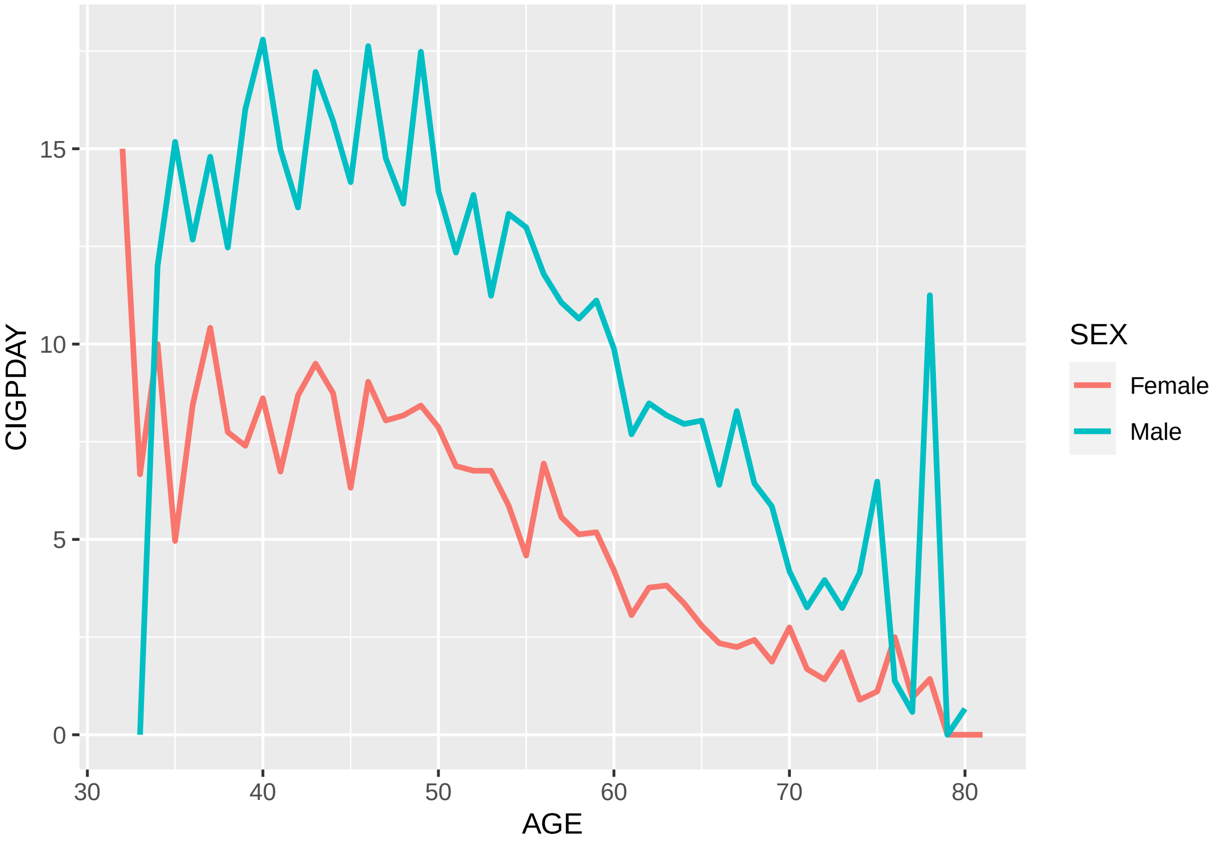
In a new plot, modify the side-by-side box-plots we created to be faceted by the sex of the participant using the facet_grid() function and columns to break up the subplots.

```
ggplot(
  data = framinghamSub,
  mapping = aes(
    x = CURSMOKE,
    y = TOTCHOL,
    fill = CURSMOKE
  )
) + geom_boxplot() + facet_grid(. ~ SEX) + scale_fill_viridis_d() + labs(
  title = "Total cholesterol by smoking status" ,
  x = "Current smoker" ,
  y = "Serum Total Cholesterol (mg/dL)",
  caption = "Data source: Framingham Heart Study & the riskCommunicator package "
)   + theme_economist() +
  theme(
    text = element_text(face = "bold"), legend.position = "none",  axis.title =
     ↪  element_text(size = 14))
```



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).

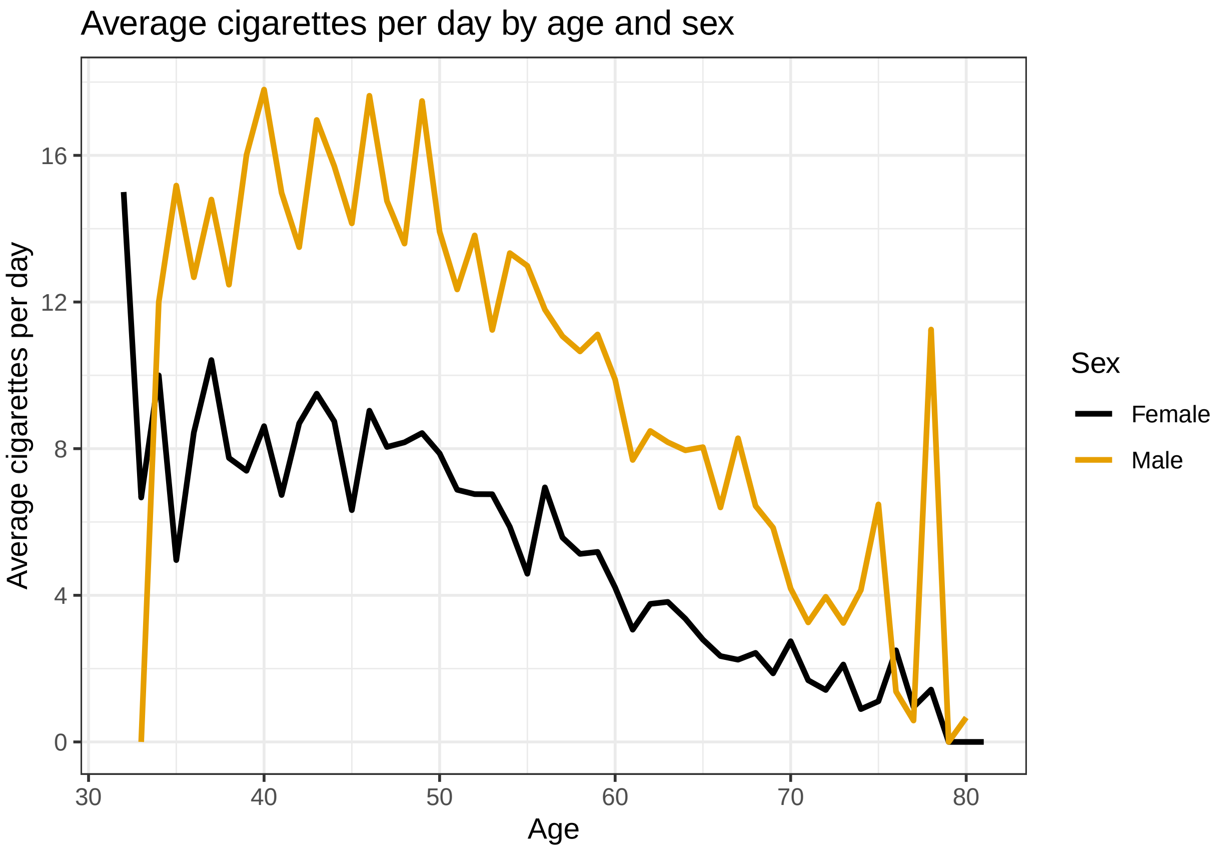
```
framinghamSub %>% ggplot() + stat_summary(aes(x= AGE, y= CIGPDAY, group = SEX, color =
↪  SEX), geom = "line", size = 1, fun.y = mean)
```



Apply a complete theme to the plot, and have the axis show the breaks at 0, 4, 8, 12, and 16 cigarettes per day.

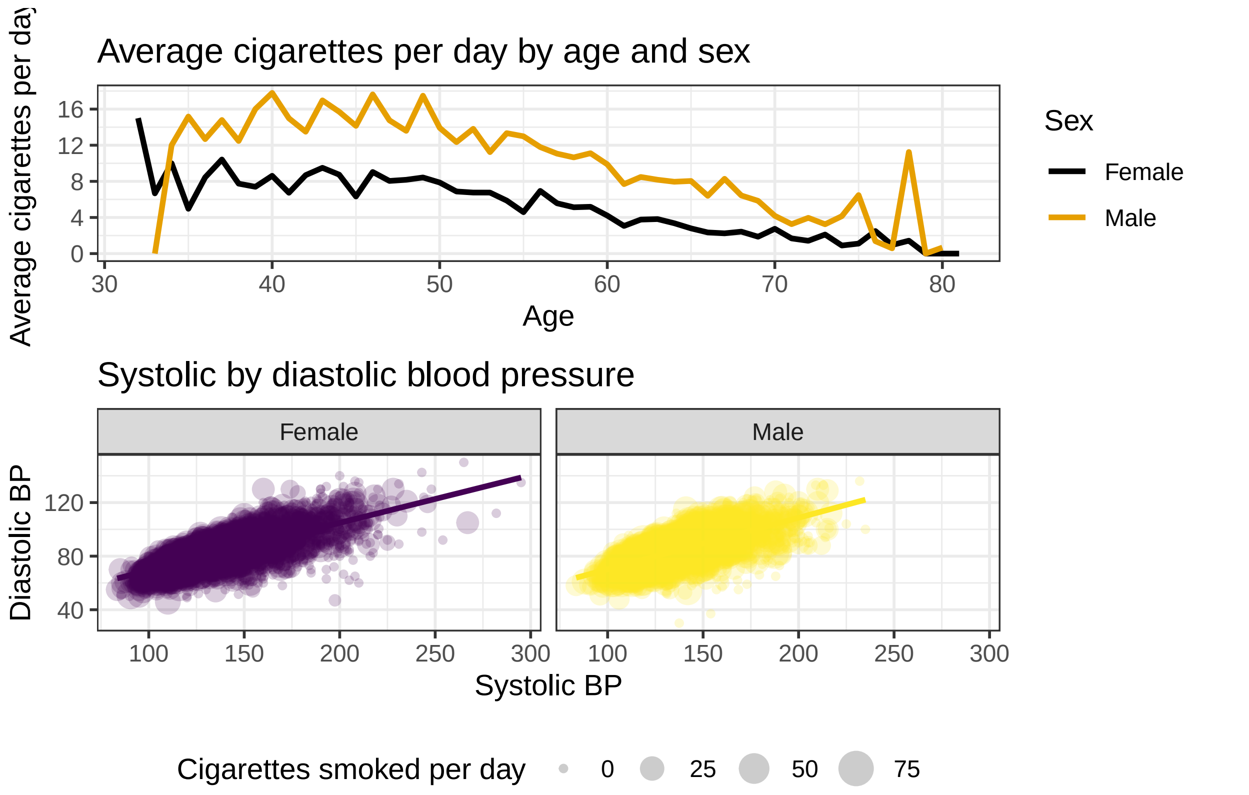
```
lineChart <- framinghamSub %>% ggplot() + stat_summary(aes(x= AGE, y= CIGPDAY, group =
 ↪  SEX, color = SEX), geom = "line", size = 1, fun.y = mean) +
  labs(title = "Average cigarettes per day by age and sex" ,
  x = "Age" ,
  y = "Average cigarettes per day",
  color = "Sex") + scale_color_colorblind() + scale_y_continuous(breaks=c(0,4,8,12,16)) +
   ↪  theme_bw()

lineChart
```



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.

```
lineChart / scatter
```



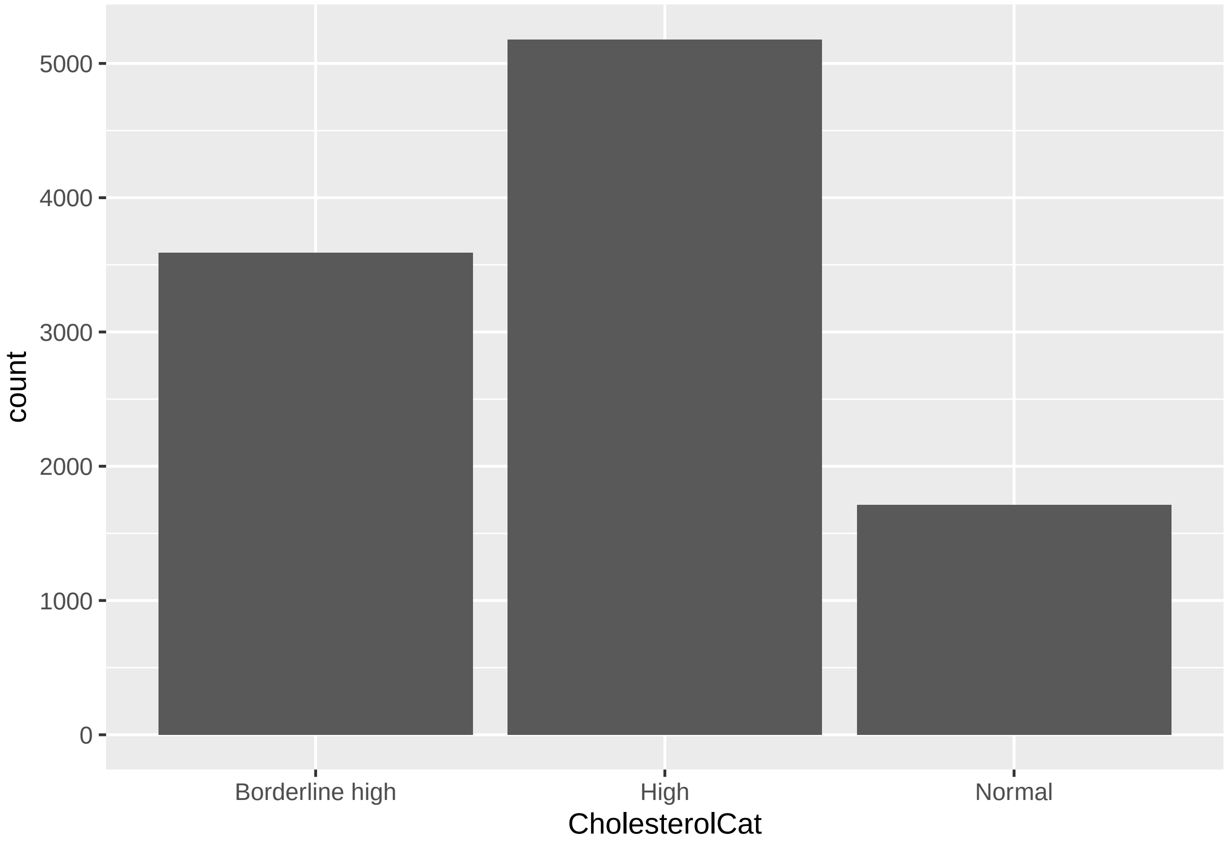
Data source: Framingham Heart Study & the riskCommunicator package

Bin / categorize total cholesterol levels is as Normal (<200 mg/dL), Borderline high (200 to 239 mg/dL), or High (> 240 mg/dL).

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

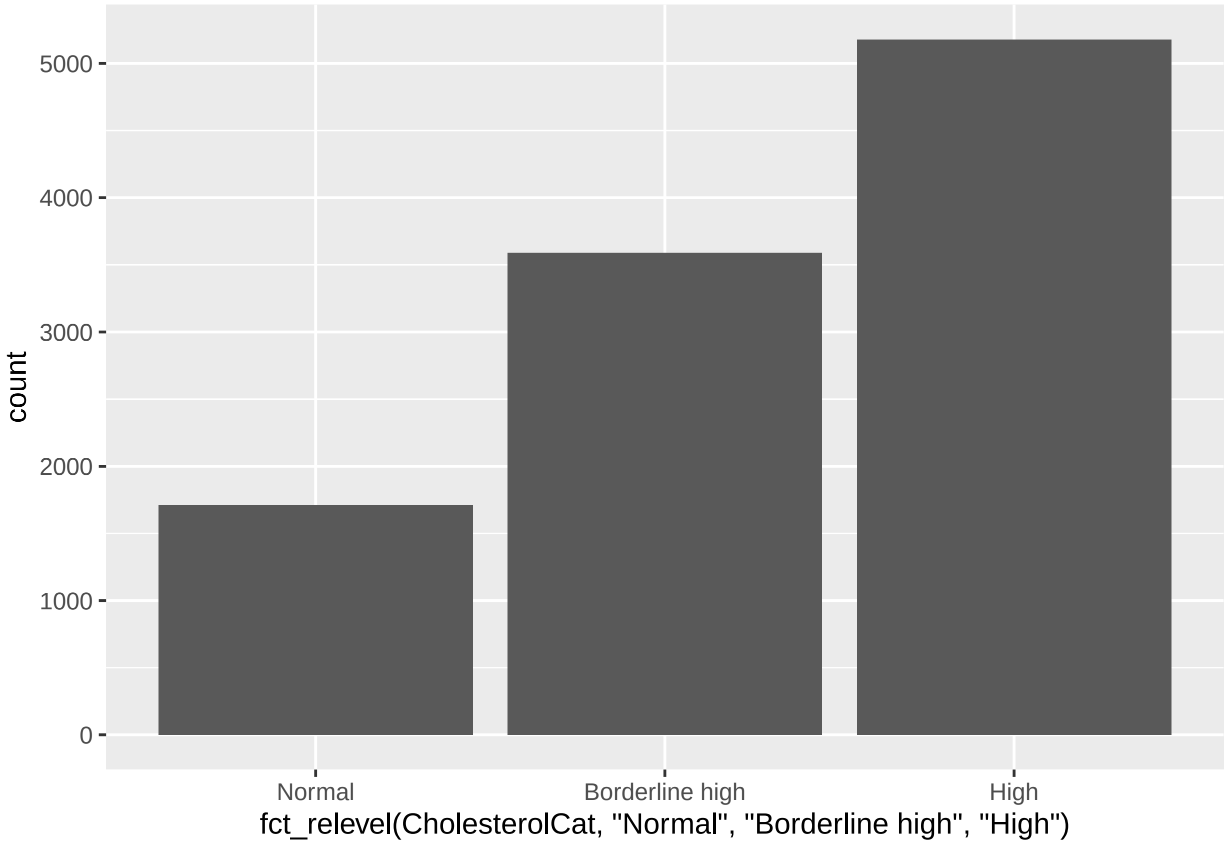
Create a bar chart displaying the number of participants falling in each cholesterol category based on Johns Hopkins' definitions using geom_bar(). Also, remove people under 40 and those without recorded cholesterol levels (missing values for CholesterolCat) from the plot by using the code filter(AGE >= 40, !is.na(CholesterolCat)) when piping the data into each subsequent ggplot() call.

```
framinghamSub %>%  filter(AGE >= 40, !is.na(CholesterolCat)) %>%
 ↪  ggplot(aes(x=CholesterolCat)) + geom_bar()
```



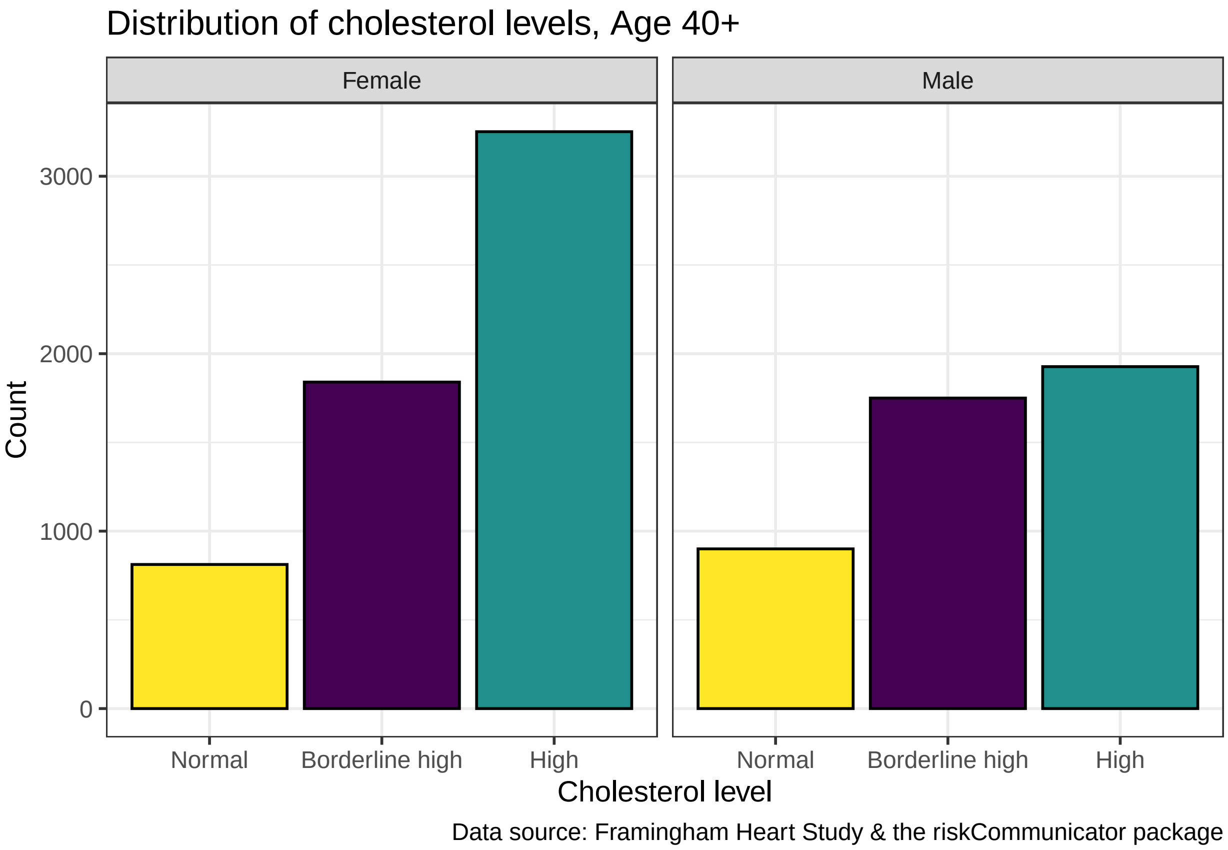
Recreate the bar chart, this time reordering the categories to show Normal, Borderline high, and then High from left to right using the fct_relevel() function.

```
framinghamSub %>%  filter(AGE >= 40, !is.na(CholesterolCat)) %>%
↪   ggplot(aes(x=fct_relevel(CholesterolCat, "Normal", "Borderline high", "High"))) +
↪   geom_bar()
```



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.

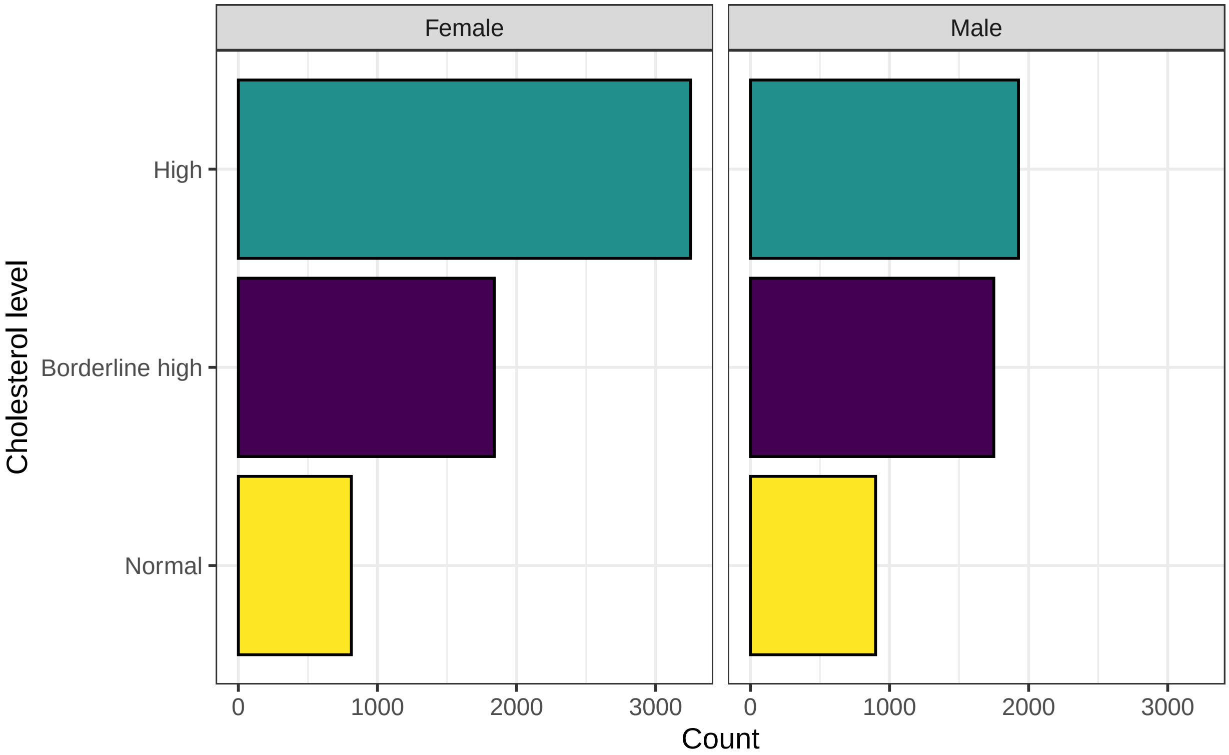
```
framinghamSub %>%  filter(AGE >= 40, !is.na(CholesterolCat)) %>%
↪  ggplot(aes(x=fct_relevel(CholesterolCat, "Normal", "Borderline high", "High"), fill =
↪  CholesterolCat ))  +
  labs(title = "Distribution of cholesterol levels, Age 40+" ,
  x = "Cholesterol level" ,
  y = "Count",
  caption = "Data source: Framingham Heart Study & the riskCommunicator package") +
   ↪  facet_grid(. ~ SEX) + scale_fill_viridis_d()+ theme_bw()  + geom_bar(color =
   ↪  "black") + theme(legend.position = "none")
```



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

```
framinghamSub %>%  filter(AGE >= 40, !is.na(CholesterolCat)) %>%
 ↪  ggplot(aes(x=fct_relevel(CholesterolCat, "Normal", "Borderline high", "High"), fill =
 ↪  CholesterolCat ))  +
  labs(title = "Distribution of cholesterol levels, Age 40+" ,
  x = "Cholesterol level" ,
  y = "Count",
  caption = "Data source: Framingham Heart Study & the riskCommunicator package") +
   ↪  facet_grid(. ~ SEX) + scale_fill_viridis_d()+ theme_bw()  + geom_bar(color =
   ↪  "black") + theme(legend.position = "none") + coord_flip()
```

Distribution of cholesterol levels, Age 40+



Data source: Framingham Heart Study & the riskCommunicator package