Intermediate-ggplot2

Arjun Mannem

2024-02-03

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

First, let’s load the FHS data set from the riskCommunicator package using the code below. Then, submit ?framingham to the Console to view the help file for the Framingham data set in RStudio.

data(framingham, package = "riskCommunicator")

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.

framinghamSub <- framingham |>  
 dplyr::select(1:10) |>  
 mutate(SexNew = case\_when(SEX == 1 ~ "Male",  
 SEX == 2 ~ "Female",  
 TRUE ~ "Other"),  
 CurSmokeNew = case\_when(CURSMOKE == 0 ~ "No",  
 CURSMOKE == 1 ~ "Yes",  
 TRUE ~ "Other"))

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

library(skimr)  
  
skim(framinghamSub)

Data summary

|  |  |
| --- | --- |
| Name | framinghamSub |
| Number of rows | 11627 |
| Number of columns | 12 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 2 |
| numeric | 10 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| SexNew | 0 | 1 | 4 | 6 | 0 | 2 | 0 |
| CurSmokeNew | 0 | 1 | 2 | 3 | 0 | 2 | 0 |

**Variable type: numeric**

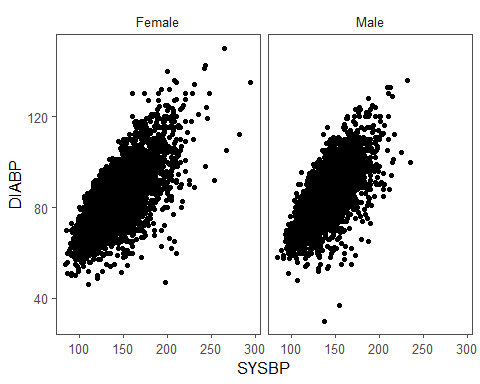
| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| RANDID | 0 | 1.00 | 5004740.92 | 2900877.44 | 2448.00 | 2474378.00 | 5006008.00 | 7472730.00 | 9999312.0 | ▇▇▇▇▇ |
| SEX | 0 | 1.00 | 1.57 | 0.50 | 1.00 | 1.00 | 2.00 | 2.00 | 2.0 | ▆▁▁▁▇ |
| 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 | ▁▅▇▁▁ |
| CURSMOKE | 0 | 1.00 | 0.43 | 0.50 | 0.00 | 0.00 | 0.00 | 1.00 | 1.0 | ▇▁▁▁▆ |
| 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 | ▇▁▁▁▁ |

Modify the code below to set a default ggplot theme for the entire document to a complete theme of your choice from the ggplot2 package or the ggthemes package.

theme\_set(ggthemes::theme\_few())

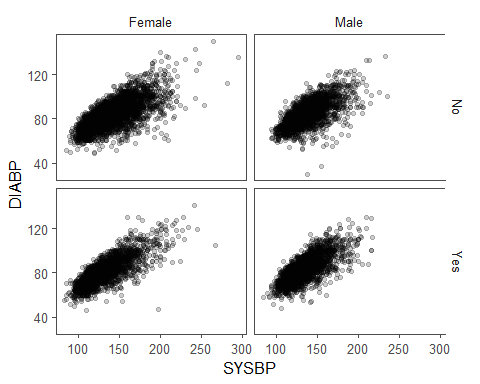
Make a scatter plot between diastolic (DIABP) and systolic (SYSBP) blood pressure with a “facet” by the sex of the participant (SEX).

framinghamSub |>   
 ggplot(aes(x = SYSBP, y = DIABP)) +  
 geom\_point() +  
 facet\_grid(. ~ SexNew) #Try facet\_grid(CurSmokeNew ~ SexNew)



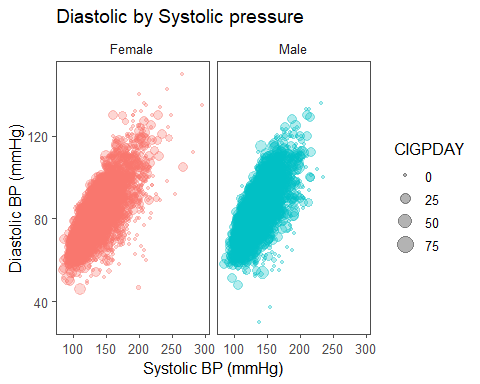
Also manually set the alpha aesthetic to be 0.2.

framinghamSub |>   
 ggplot(aes(x = SYSBP, y = DIABP)) +  
 geom\_point(alpha = 0.2) +  
 facet\_grid(CurSmokeNew ~ SexNew)



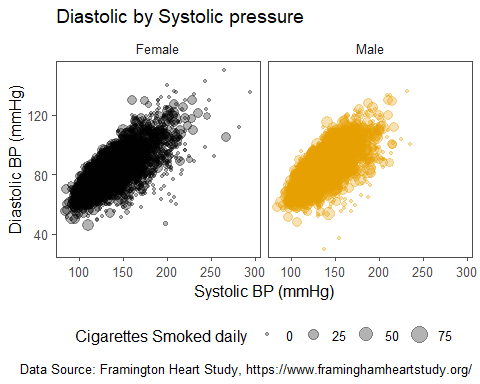
Add a guides(color = FALSE) layer to suppress the legend since it is redundant.

framinghamSub |>  
 ggplot(aes(x = SYSBP, y = DIABP,  
 size = CIGPDAY,  
 color = SexNew)) +  
 geom\_point(alpha = 0.3) +  
 facet\_grid(. ~ SexNew) +  
 guides(color = "none") +  
 labs(title = "Diastolic by Systolic pressure",  
 y = "Diastolic BP (mmHg)",  
 x = "Systolic BP (mmHg)")



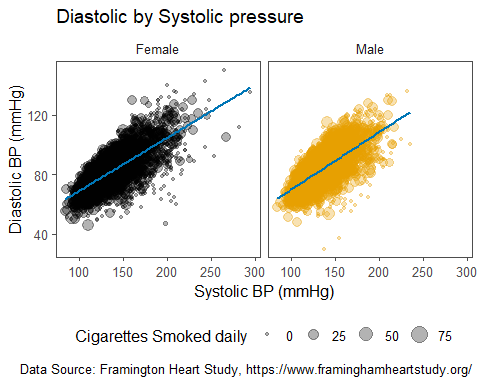
Also include the size of the data points as mapped by the number of cigarettes smoked per day (CIGPDAY), add a color-blind friendly palette for coloring the points, and position the legend at the bottom of the plot.

framinghamSub |>  
 ggplot(aes(x = SYSBP, y = DIABP,  
 size = CIGPDAY,  
 color = SexNew)) +  
 geom\_point(alpha = 0.3) +  
 scale\_color\_colorblind() +  
 facet\_grid(. ~ SexNew) +  
 guides(color = "none") +  
 labs(title = "Diastolic by Systolic pressure",  
 y = "Diastolic BP (mmHg)",  
 x = "Systolic BP (mmHg)",  
 caption = "Data Source: Framington Heart Study, https://www.framinghamheartstudy.org/",  
 size = "Cigarettes Smoked daily") +  
 theme(legend.position = "bottom")



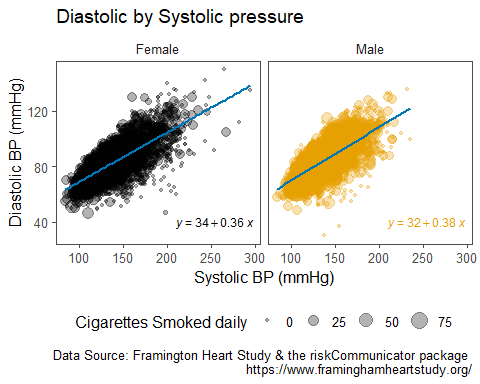
Add a line of best fit corresponding to a simple linear regression model fit separately for males and females using geom\_smooth().

framinghamSub |>  
 ggplot(aes(x = SYSBP, y = DIABP,  
 size = CIGPDAY,  
 color = SexNew)) +  
 geom\_point(alpha = 0.3) +  
 geom\_smooth(method = "lm", se = FALSE,  
 color = "#0076B6", size = 0.8) +  
 scale\_color\_colorblind() +  
 facet\_grid(. ~ SexNew) +  
 guides(color = "none") +  
 labs(title = "Diastolic by Systolic pressure",  
 y = "Diastolic BP (mmHg)",  
 x = "Systolic BP (mmHg)",  
 caption = "Data Source: Framington Heart Study, https://www.framinghamheartstudy.org/",  
 size = "Cigarettes Smoked daily") +  
 theme(legend.position = "bottom")



Add the estimated regression equations to each sub-plot using the ggpubr package and adding a stat\_regline\_equation(label.x = 210, label.y = 40, size = 3.2) layer.

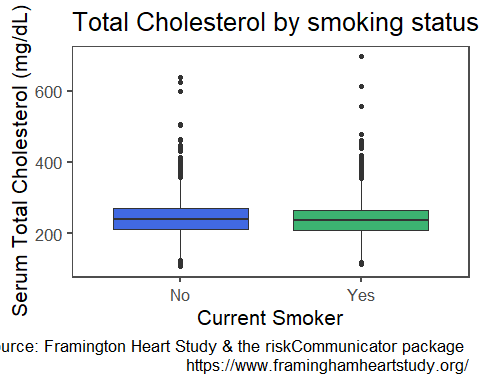
framinghamSub |>  
 ggplot(aes(x = SYSBP, y = DIABP,  
 size = CIGPDAY,  
 color = SexNew)) +  
 geom\_point(alpha = 0.3) +  
 geom\_smooth(method = "lm", se = FALSE,  
 color = "#0076B6", size = 0.8) +  
 scale\_color\_colorblind() +  
 facet\_grid(. ~ SexNew) +  
 guides(color = "none") +  
 labs(title = "Diastolic by Systolic pressure",  
 y = "Diastolic BP (mmHg)",  
 x = "Systolic BP (mmHg)",  
 caption = "Data Source: Framington Heart Study & the riskCommunicator package \n https://www.framinghamheartstudy.org/",  
 size = "Cigarettes Smoked daily") +  
 stat\_regline\_equation(label.x = 210,label.y = 40,size = 3.2) +  
 theme(legend.position = "bottom")



## Boxplots

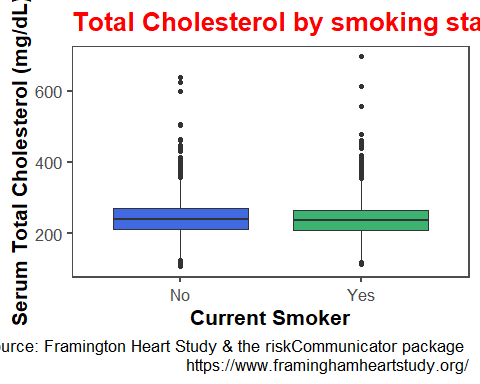
Next, create a side-by-side box-plot where the y-axis is total cholesterol (TOTCHOL) and the x-axis is current smoking status (CURSMOKE). Increase the font size for the axes and title text in the plot

# creating a side-by-side box plots  
framinghamSub |>  
 ggplot(aes(x = CurSmokeNew, y = TOTCHOL, fill = CurSmokeNew)) +  
 geom\_boxplot() +  
 scale\_fill\_manual(values = c("royalblue","mediumseagreen")) +  
 labs(title = "Total Cholesterol by smoking status",  
 y = "Serum Total Cholesterol (mg/dL)",  
 x = "Current Smoker",  
 caption = "Data Source: Framington Heart Study & the riskCommunicator package \n https://www.framinghamheartstudy.org/") +  
 theme\_few(base\_size = 16) +  
 theme(legend.position = "none")



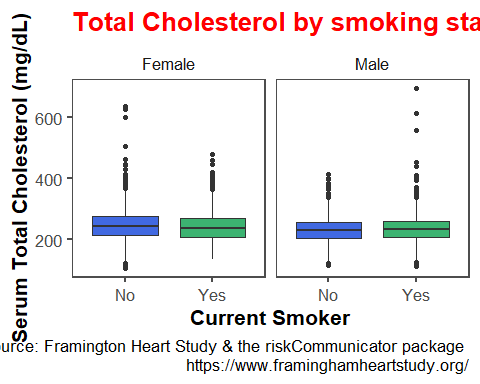
Color the boxes based on smoking status by manually specifying the colors to be “mediumseagreen” and “royalblue”, remove the legend, and make the title and axis titles bold.

# creating a side-by-side box plots  
framinghamSub |>  
 ggplot(aes(x = CurSmokeNew, y = TOTCHOL, fill = CurSmokeNew)) +  
 geom\_boxplot() +  
 scale\_fill\_manual(values = c("royalblue","mediumseagreen")) +  
 labs(title = "Total Cholesterol by smoking status",  
 y = "Serum Total Cholesterol (mg/dL)",  
 x = "Current Smoker",  
 caption = "Data Source: Framington Heart Study & the riskCommunicator package \n https://www.framinghamheartstudy.org/") +  
 theme\_few(base\_size = 16) +  
 theme(legend.position = "none",   
 plot.title = element\_text(face = "bold", color = "red"),  
 axis.title.x = element\_text(face = "bold"),  
 axis.title.y = element\_text(face = "bold"))



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.

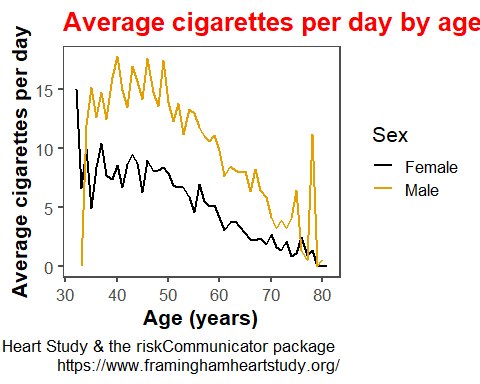
# creating a side-by-side box plots  
framinghamSub |>  
 ggplot(aes(x = CurSmokeNew, y = TOTCHOL, fill = CurSmokeNew)) +  
 geom\_boxplot() +  
 facet\_grid(. ~ SexNew) +  
 scale\_fill\_manual(values = c("royalblue","mediumseagreen")) +  
 labs(title = "Total Cholesterol by smoking status",  
 y = "Serum Total Cholesterol (mg/dL)",  
 x = "Current Smoker",  
 caption = "Data Source: Framington Heart Study & the riskCommunicator package \n https://www.framinghamheartstudy.org/") +  
 theme\_few(base\_size = 16) +  
 theme(legend.position = "none",   
 plot.title = element\_text(face = "bold", color = "red"),  
 axis.title.x = element\_text(face = "bold"),  
 axis.title.y = element\_text(face = "bold"))



## Linechart

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).

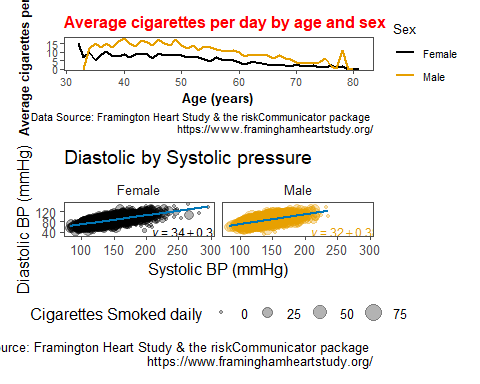
framinghamSub |>  
 ggplot() +  
 stat\_summary(aes(x = AGE,  
 y = CIGPDAY,  
 color = SexNew,  
 group = SexNew),  
 geom = "line",  
 fun.y = mean,  
 size = 1) +  
 scale\_color\_colorblind() +  
 labs(title = "Average cigarettes per day by age and sex",  
 y = "Average cigarettes per day",  
 x = "Age (years)",  
 color = "Sex",  
 caption = "Data Source: Framington Heart Study & the riskCommunicator package \n https://www.framinghamheartstudy.org/") +  
 theme\_few(base\_size = 16) +  
 theme(plot.title = element\_text(face = "bold", color = "red"),  
 axis.title.x = element\_text(face = "bold"),  
 axis.title.y = element\_text(face = "bold"))



## Combining plots with patchwork

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.

my\_linechart <- framinghamSub |>  
 ggplot() +  
 stat\_summary(aes(x = AGE,  
 y = CIGPDAY,  
 color = SexNew,  
 group = SexNew),  
 geom = "line",  
 fun.y = mean,  
 size = 1) +  
 scale\_color\_colorblind() +  
 labs(title = "Average cigarettes per day by age and sex",  
 y = "Average cigarettes per day",  
 x = "Age (years)",  
 color = "Sex",  
 caption = "Data Source: Framington Heart Study & the riskCommunicator package \n https://www.framinghamheartstudy.org/") +  
 theme\_few(base\_size = 10) +  
 theme(plot.title = element\_text(face = "bold", color = "red"),  
 axis.title.x = element\_text(face = "bold"),  
 axis.title.y = element\_text(face = "bold"))  
  
my\_scatterplot <- framinghamSub |>  
 ggplot(aes(x = SYSBP, y = DIABP,  
 size = CIGPDAY,  
 color = SexNew)) +  
 geom\_point(alpha = 0.3) +  
 geom\_smooth(method = "lm", se = FALSE,  
 color = "#0076B6", size = 0.8) +  
 scale\_color\_colorblind() +  
 facet\_grid(. ~ SexNew) +  
 guides(color = "none") +  
 labs(title = "Diastolic by Systolic pressure",  
 y = "Diastolic BP (mmHg)",  
 x = "Systolic BP (mmHg)",  
 caption = "Data Source: Framington Heart Study & the riskCommunicator package \n https://www.framinghamheartstudy.org/",  
 size = "Cigarettes Smoked daily") +  
 stat\_regline\_equation(label.x = 210,label.y = 40,size = 3.2) +  
 theme(legend.position = "bottom")  
  
library(patchwork)  
  
my\_linechart / my\_scatterplot



## Barchart

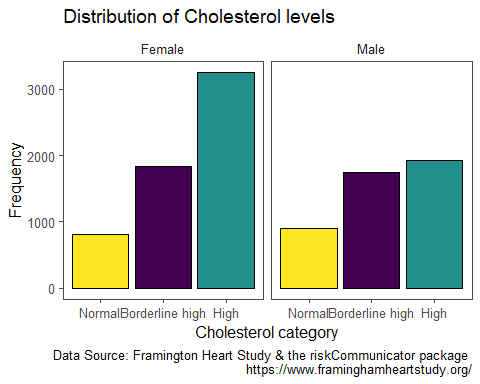
framinghamSub <- framinghamSub |>   
 mutate(CholesterolCat = case\_when(TOTCHOL < 200 ~ "Normal",  
 TOTCHOL >= 200 & TOTCHOL < 240 ~ "Borderline high",  
 TOTCHOL > 240 ~ "High",  
 TRUE ~ as.character(NA)))

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.

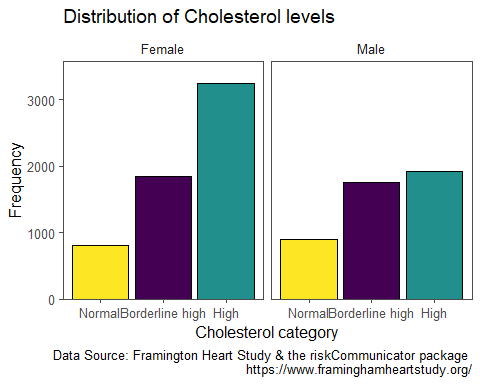
Change the color of the inside of the bars based on the cholesterol category using a color-blind friendly palette, make the outline of the bars black in color, facet by the sex of the participant, and remove the legend.

my\_bar <- framinghamSub |>   
 filter(AGE >= 40, !is.na(CholesterolCat)) |>  
 ggplot(aes(x = fct\_relevel(CholesterolCat,"Normal","Borderline high", "High"),  
 fill = CholesterolCat)) +  
 scale\_fill\_viridis\_d() +  
 geom\_bar(color = "black") +   
 facet\_grid(.~SexNew) +  
 labs(title = "Distribution of Cholesterol levels",  
 x = "Cholesterol category",  
 y = "Frequency",  
 caption = "Data Source: Framington Heart Study & the riskCommunicator package \n https://www.framinghamheartstudy.org/") +  
 theme(legend.position = "none")  
  
my\_bar



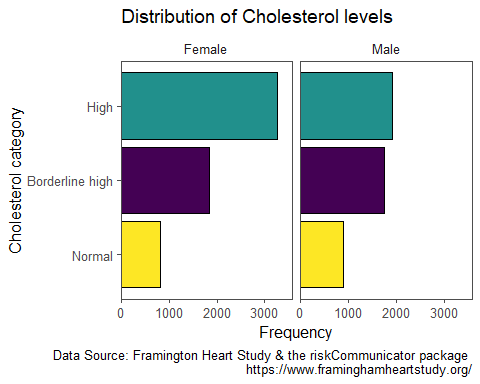
Remove the extra space between the bars and the horizontal axis.

my\_bar <- my\_bar +  
 scale\_y\_continuous(expand = expansion(mult = c(0,0.1)))  
  
my\_bar



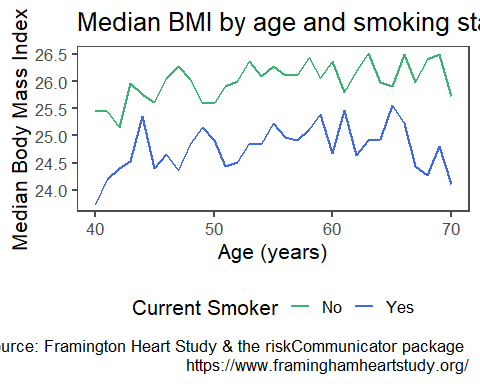
Lastly, use the coor\_flip() function to turn the bar chart into a horizontal bar chart instead.

my\_bar +  
 coord\_flip()



Recreate the line chart below showing the median BMI by age for smokers and non-smokers. Hint: use a scale\_color\_manual() layer to match the colors.

framinghamSub |>  
 ggplot() +  
 stat\_summary(aes(x = AGE,  
 y = BMI,  
 color = CurSmokeNew,  
 group = CurSmokeNew),  
 geom = "line",  
 fun.y = median,  
 size = 1) +  
 scale\_color\_manual(values = c("mediumseagreen","royalblue")) +  
 labs(title = "Median BMI by age and smoking status",  
 y = "Median Body Mass Index",  
 x = "Age (years)",  
 color = "Current Smoker",  
 caption = "Data Source: Framington Heart Study & the riskCommunicator package \n https://www.framinghamheartstudy.org/") +  
 theme\_few(base\_size = 16) +  
 theme(legend.position = "bottom") +  
 xlim(40,70)



We can use scale\_x\_continuous(limits = c(40,70)) or xlim(40,70) or dplyr::filter(between(AGE.40,70))

Recreate the faceted dumbbell chart below showing the same information as the faceted bar chart. Hint: One way to create a dumbbell chart is to use geom\_point(size = 4, pch = 21) and geom\_line().

framinghamSub |>   
 dplyr::count(CholesterolCat,SexNew,.drop = FALSE) |>  
 dplyr::filter( !is.na(CholesterolCat)) |>  
 ggplot(aes(x = n,y = SexNew, color = CholesterolCat, fill = CholesterolCat)) +   
 scale\_fill\_viridis\_d() +  
 geom\_line(aes(group = SexNew),color = "black") +  
 geom\_point(color = "black",size = 4, pch = 21) +  
 labs(title = "Distribution of Cholesterol levels",  
 x = "Frequency",  
 y = "",  
 fill = "Cholesterol category:",  
 caption = "Data Source: Framington Heart Study & the riskCommunicator package \n https://www.framinghamheartstudy.org/") +  
 theme(legend.position = "bottom")

