Module 2

## What is Quarto?

Quarto enables you to weave together content and executable code into a finished document. To learn more about Quarto see <https://quarto.org>.

## Quarto Basics- Running Code

When you click the **Render** button a document will be generated that includes both content and the output of embedded code. You can embed code like this:

1 + 1

[1] 2

You can add options to executable code like this

[1] 4

The echo: false option disables the printing of code (only output is displayed).

## Frequency Tables

Let’s create a frequency table for the variable ChestPainType within the heart data set. Don’t forget to read in the data first.

heart <- read.csv("C:/Users/jenni/Dropbox (ASU)/Courses/LSC540/Data Sets/heart.csv")   
my.table <- table(heart$ChestPainType, dnn="Chest Pain Type")  
my.table

Chest Pain Type  
ASY ATA NAP TA   
496 173 203 46

Let’s add the total to the table.

addmargins(my.table)

Chest Pain Type  
ASY ATA NAP TA Sum   
496 173 203 46 918

Let’s create a relative frequency table.

my.table2 <- table(heart$HeartDisease, dnn=c("Heart Disease (1=yes, 0=no)"))  
prop.table(my.table2)

Heart Disease (1=yes, 0=no)  
 0 1   
0.4466231 0.5533769

round(prop.table(my.table2), 4) # Rounded

Heart Disease (1=yes, 0=no)  
 0 1   
0.4466 0.5534

Create a pretty table that is more visually appealing.

knitr::kable(my.table)

| Chest.Pain.Type | Freq |
| --- | --- |
| ASY | 496 |
| ATA | 173 |
| NAP | 203 |
| TA | 46 |

## Contingency Tables

my.ctable <- table(heart$ChestPainType, heart$HeartDisease,dnn=c("Heart Disease (1=yes, 0=no)","Chest Pain Type" ))  
my.ctable

Chest Pain Type  
Heart Disease (1=yes, 0=no) 0 1  
 ASY 104 392  
 ATA 149 24  
 NAP 131 72  
 TA 26 20

knitr::kable(my.ctable)

|  | 0 | 1 |
| --- | --- | --- |
| ASY | 104 | 392 |
| ATA | 149 | 24 |
| NAP | 131 | 72 |
| TA | 26 | 20 |

addmargins(my.ctable) # Add Row and Column Totals

Chest Pain Type  
Heart Disease (1=yes, 0=no) 0 1 Sum  
 ASY 104 392 496  
 ATA 149 24 173  
 NAP 131 72 203  
 TA 26 20 46  
 Sum 410 508 918

prop.table(my.ctable)

Chest Pain Type  
Heart Disease (1=yes, 0=no) 0 1  
 ASY 0.11328976 0.42701525  
 ATA 0.16230937 0.02614379  
 NAP 0.14270153 0.07843137  
 TA 0.02832244 0.02178649

## Bar Graphs

Now let’s visualize the summary data

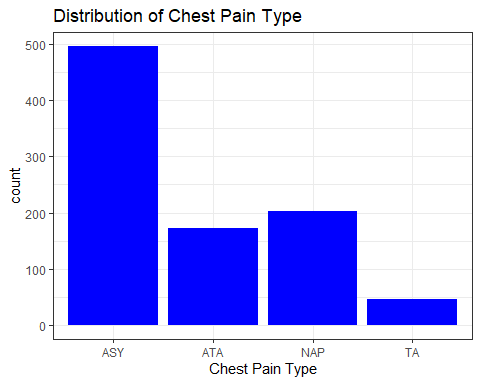
library(ggplot2) # Loads the ggplot2 library  
library(dplyr) # Loads the dplyr library

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':  
  
 filter, lag

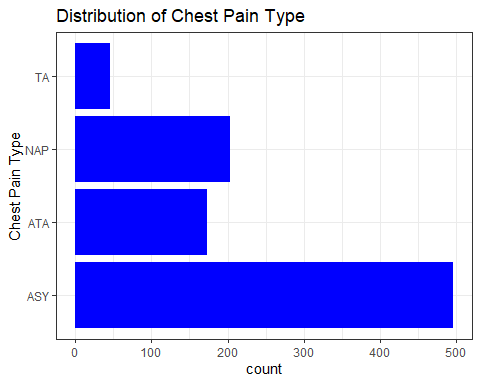
The following objects are masked from 'package:base':  
  
 intersect, setdiff, setequal, union

heart %>%   
 ggplot( aes(x = ChestPainType)) +   
 geom\_bar(fill = "blue") +   
 labs(  
 title = "Distribution of Chest Pain Type",   
 x = "Chest Pain Type"  
 ) +   
 theme\_bw()



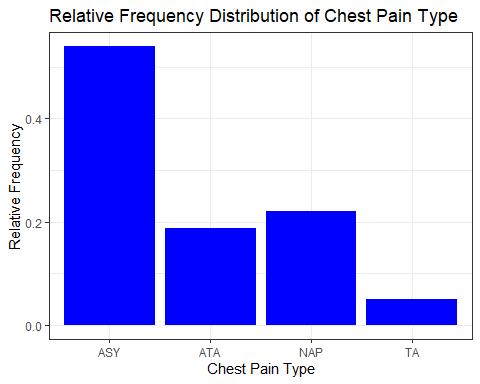
Flip to horizontal bars:

heart %>%   
 ggplot( aes(x = ChestPainType)) +   
 geom\_bar(fill = "blue") +   
 labs(  
 title = "Distribution of Chest Pain Type",   
 x = "Chest Pain Type"  
 ) +   
 theme\_bw() +   
 coord\_flip()



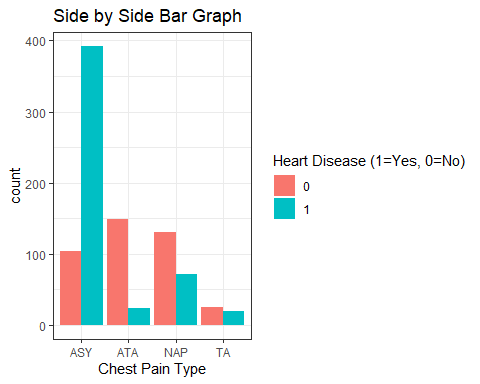
Change to relative frequency

heart %>%   
 ggplot( aes(x = ChestPainType, y = after\_stat(count / sum(count)))) +   
 geom\_bar(fill = "blue") +   
 labs(  
 title = "Relative Frequency Distribution of Chest Pain Type",   
 x = "Chest Pain Type",   
 y= "Relative Frequency"  
 ) +   
 theme\_bw()

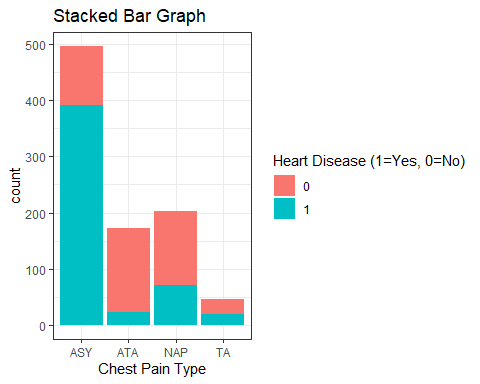


## Stacked and Grouped Bar Graphs

heart %>%   
 ggplot( aes(x = ChestPainType, fill =as.factor(HeartDisease) )) +   
 geom\_bar(position="dodge") +   
 labs(  
 title = "Side by Side Bar Graph",   
 x = "Chest Pain Type",   
 fill= "Heart Disease (1=Yes, 0=No)"  
 ) +   
 theme\_bw()



heart %>%   
 ggplot( aes(x = ChestPainType, fill =as.factor(HeartDisease) )) +   
 geom\_bar() +   
 labs(  
 title = "Stacked Bar Graph",   
 x = "Chest Pain Type",   
 fill= "Heart Disease (1=Yes, 0=No)"  
 ) +   
 theme\_bw()



# Module 2.2

## Univariate Summary statistics

#Measures of center  
mean(heart$Age)

[1] 53.51089

median(heart$Age)

[1] 54

#Measures of spread  
sd(heart$Age)

[1] 9.432617

var(heart$Age)

[1] 88.97425

diff(range(heart$Age))

[1] 49

range(heart$Age)

[1] 28 77

summary(heart$Age)

Min. 1st Qu. Median Mean 3rd Qu. Max.   
 28.00 47.00 54.00 53.51 60.00 77.00

## Comparing Across Groups

library(mosaic) # Loads the mosaic package

Registered S3 method overwritten by 'mosaic':  
 method from   
 fortify.SpatialPolygonsDataFrame ggplot2

The 'mosaic' package masks several functions from core packages in order to add   
additional features. The original behavior of these functions should not be affected by this.

Attaching package: 'mosaic'

The following object is masked from 'package:Matrix':  
  
 mean

The following objects are masked from 'package:dplyr':  
  
 count, do, tally

The following object is masked from 'package:ggplot2':  
  
 stat

The following objects are masked from 'package:stats':  
  
 binom.test, cor, cor.test, cov, fivenum, IQR, median, prop.test,  
 quantile, sd, t.test, var

The following objects are masked from 'package:base':  
  
 max, mean, min, prod, range, sample, sum

favstats(Age ~ HeartDisease, data = heart)

HeartDisease min Q1 median Q3 max mean sd n missing  
1 0 28 43 51 57 76 50.55122 9.444915 410 0  
2 1 31 51 57 62 77 55.89961 8.727056 508 0

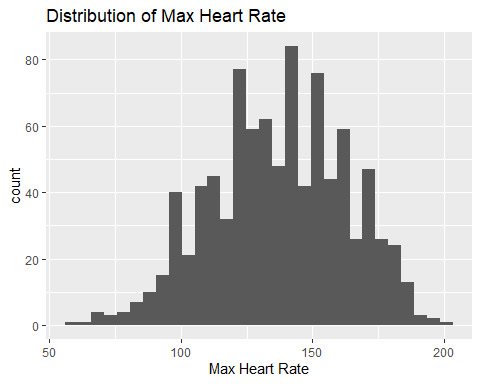
favstats(MaxHR ~ HeartDisease, data = heart)

HeartDisease min Q1 median Q3 max mean sd n missing  
1 0 69 134 150 165.00 202 148.1512 23.28807 410 0  
2 1 60 112 126 144.25 195 127.6555 23.38692 508 0

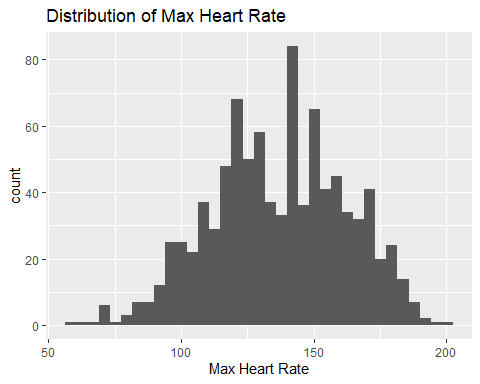
## Histograms - Univariate Numerical Data

ggplot(data = heart, aes(x = MaxHR)) +   
 geom\_histogram() +   
 ggtitle("Distribution of Max Heart Rate") +  
 xlab("Max Heart Rate")

`stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

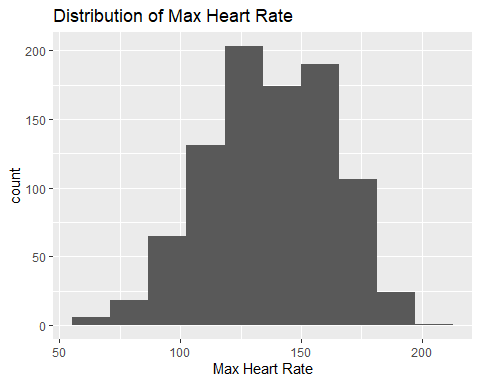


ggplot(data = heart, aes(x = MaxHR)) +   
 geom\_histogram(bins=35) +   
 ggtitle("Distribution of Max Heart Rate") +  
 xlab("Max Heart Rate")



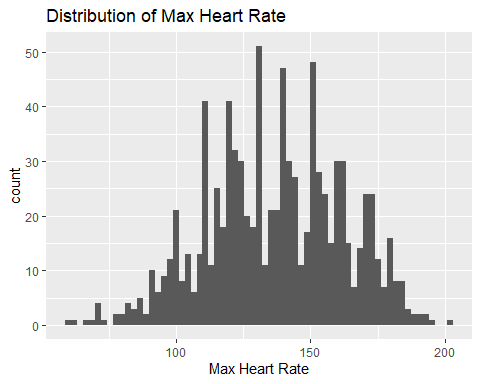
Too few bins

ggplot(data = heart, aes(x = MaxHR)) +   
 geom\_histogram(bins=10) +   
 ggtitle("Distribution of Max Heart Rate") +  
 xlab("Max Heart Rate")



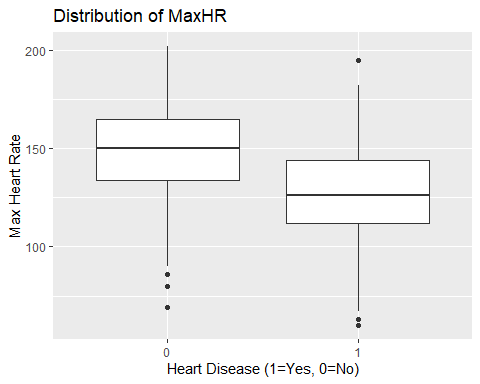
Too many bins

ggplot(data = heart, aes(x = MaxHR)) +   
 geom\_histogram(bins=65) +   
 ggtitle("Distribution of Max Heart Rate") +  
 xlab("Max Heart Rate")



## Boxplots- Comparing a numerical variable across groups

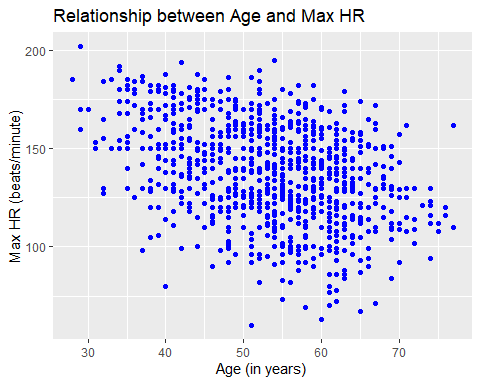
ggplot(heart, aes(x=as.factor(HeartDisease), y=MaxHR)) + geom\_boxplot() +  
 ggtitle("Distribution of MaxHR ") +  
 ylab("Max Heart Rate") +   
 xlab("Heart Disease (1=Yes, 0=No)")



# Module 2.3 - Bivariate data

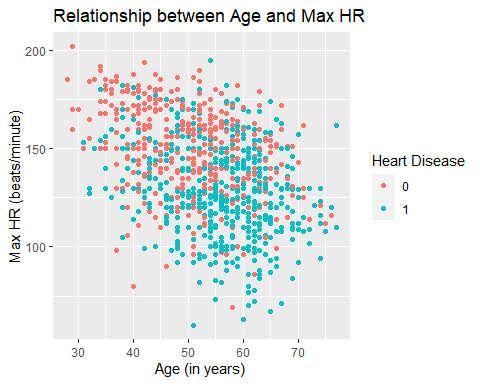
## Scatterplots

my.plot <- ggplot(data = heart, aes(x = Age, y = MaxHR)) + geom\_point(color = "blue") +   
 ggtitle("Relationship between Age and Max HR") +  
 xlab("Age (in years) ") +   
 ylab("Max HR (beats/minute)")   
my.plot



Multivariate Scatterplot - add third categorical variable

my.plot2 <- ggplot(data = heart, aes(x = Age, y = MaxHR)) + geom\_point(color = "blue") +   
 geom\_point(aes(color = factor(HeartDisease)) ) +   
 ggtitle("Relationship between Age and Max HR") +  
 xlab("Age (in years) ") +   
 ylab("Max HR (beats/minute)") +  
 scale\_color\_discrete(name="Heart Disease")  
my.plot2



Correlation coefficient

cor(heart$Age, heart$MaxHR)

[1] -0.3820447