Introduction to Data Science, Topic 4

- Instructor: Professor Henry Horng-Shing Lu,

  Institute of Statistics, National Chiao Tung University, Taiwan

  Email: hslu@stat.nctu.edu.tw

- WWW: http://www.stat.nctu.edu.tw/misg/hslu/course/DataScience.htm
- Reference:

  M. A. Pathak, Beginning Data Science with R, 2014, Springer-Verlag.
- Evaluation: Homework: 50%, Term Project: 50%
- Office hours: By appointment

Course Outline

- **Introduction of data science**
- **Introduction of R**
- **More on R**
- **Data Visualization**
- **Exploratory Data Analysis**
- **Regression**
- **Classification**
- **Text Mining**
- **Clustering**

Data Visualization with R

References:

Ch. 4, M. A. Pathak, Beginning Data Science with R, 2014, Springer-Verlag.

Its eyes are round, with a triangular mouth under the little nose, a beautiful "eight" character on either side of the mouth, and two pointed ears erected on the round head, making it particularly airy.

https://en.wiktionary.org/wiki/cat

“a picture is worth a thousand words”

Basic Visualizations

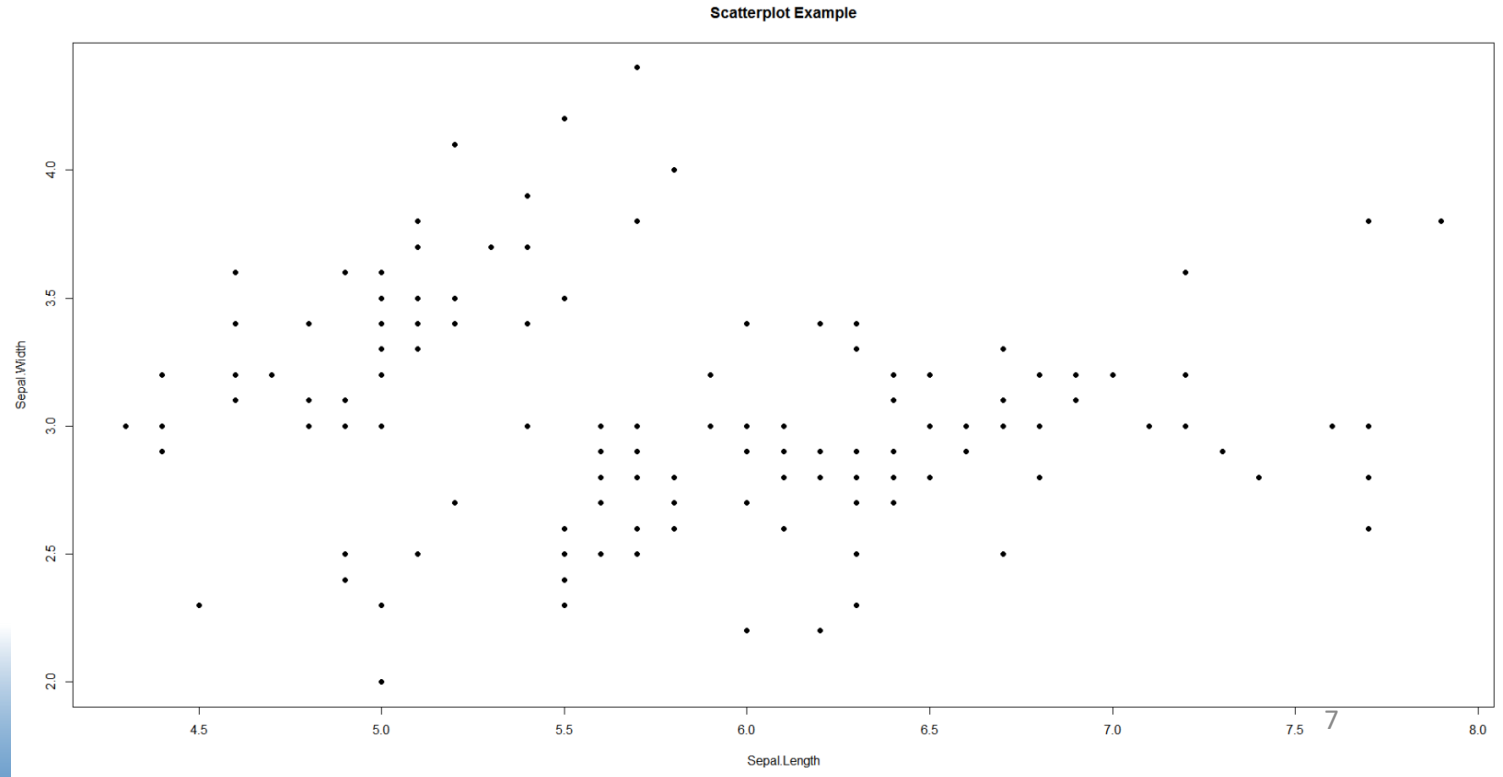
- ***Scatterplots***
- ***Visualizing Aggregate Values with Bar plots and Pie charts***
- ***Common Plotting Tasks***

Data

- Iris Data Set: (iris)
  - Sepal Length (mm)
  - Sepal Width (mm)
  - Petal Length (mm)
  - Petal Width (mm)
  - Species

Scatterplots

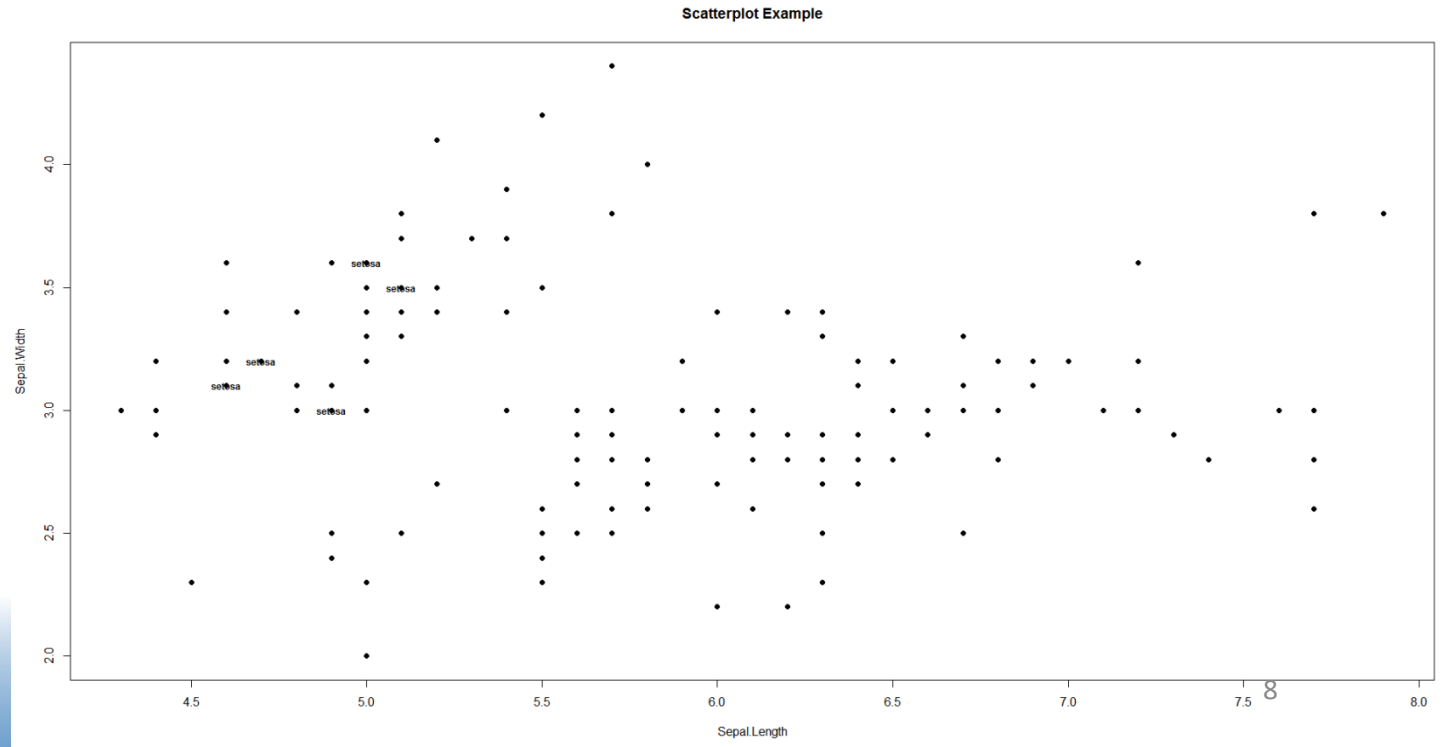
- **Basic setup**
  - The value of data
- **Code:**
  - attach(iris);
  - plot(Sepal.Length,Sepal.Width,main="Scatterplot Example",xlab="Sepal.Length",ylab="Sepal.Width", pch=19)



Scatterplots

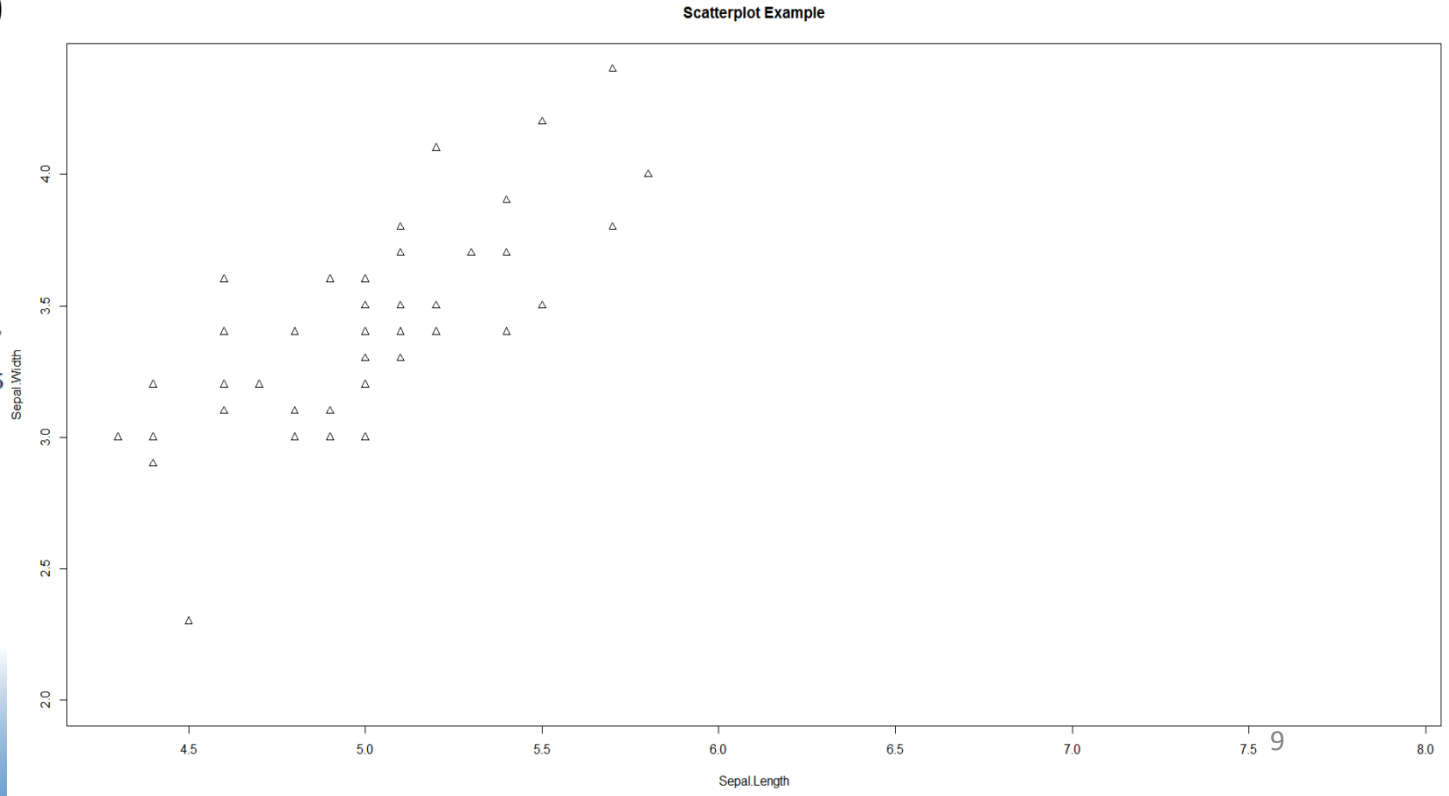
- **Labeling Data Points**

- attach(iris);
- plot(Sepal.Length,Sepal.Width,main="Scatterplot Example",xlab="Sepal.Length",ylab="Sepal.Width", pch=19)
- text(Sepal.Length[1:5],Sepal.Width[1:5], labels = Species[1:5],cex=0.75,font=2)



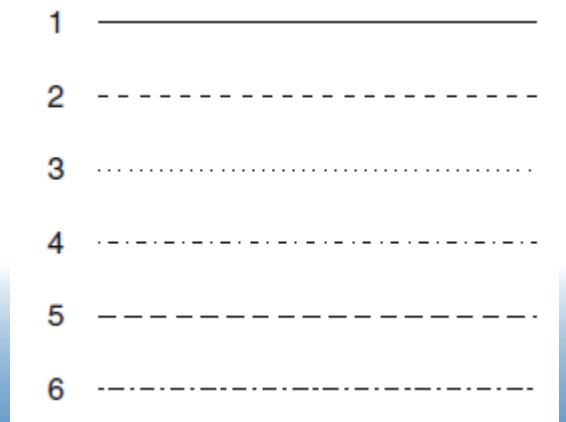
Scatterplots

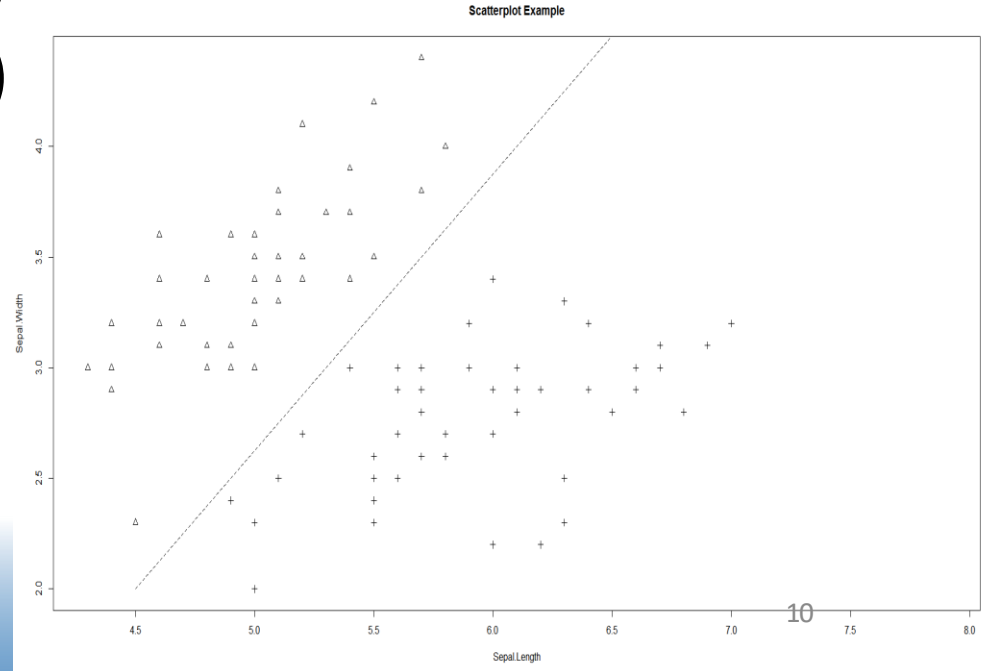
- **Points and Lines**
  - s1 = which(Species=="setosa")
  - plot(Sepal.Length,Sepal.Width,main="Scatterplot Example",xlab="Sepal.Length",ylab="Sepal.Width",type="n")
  - points(Sepal.Length[s1],Sepal.Width[s1],pch=2)



Scatterplots

- **Points and Lines**
- s1 = which(Species=="setosa");s2 = which(Species=="versicolor");
- plot(Sepal.Length,Sepal.Width,main="Scatterplot Example",xlab="Sepal.Length",ylab="Sepal.Width",type="n")
- points(Sepal.Length[s1],Sepal.Width[s1],pch=2)
- points(Sepal.Length[s2],Sepal.Width[s2],pch=3)
- lines(c(4.5,6.5),c(2,4.5),lty=2)

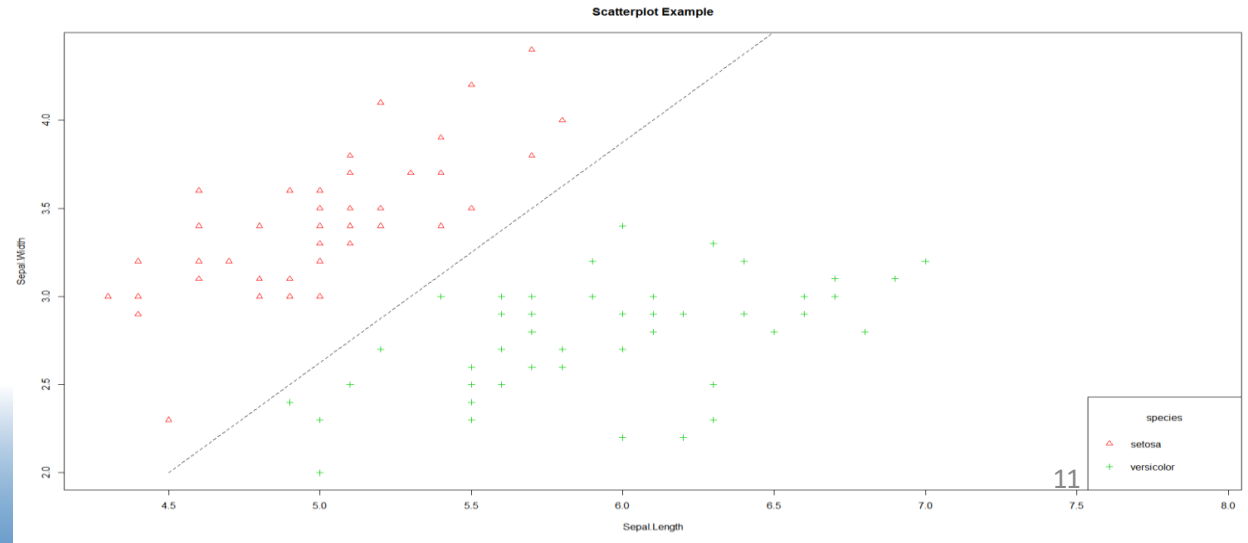




Scatterplots

- **Points and Lines**

- s1 = which(Species=="setosa");s2 = which(Species=="versicolor");
- plot(Sepal.Length,Sepal.Width,main="Scatterplot Example",xlab="Sepal.Length",ylab="Sepal.Width",type="n")
- points(Sepal.Length[s1],Sepal.Width[s1],pch=2,col = 2)
- points(Sepal.Length[s2],Sepal.Width[s2],pch=3,col=3)
- lines(c(4.5,6.5),c(2,4.5),lty=2)
- legend("bottomright",c("setosa", "versicolor"), col=c(2,3),pch=c(2,3),title="species")



Visualizing Aggregate Values with Bar plots and Pie charts

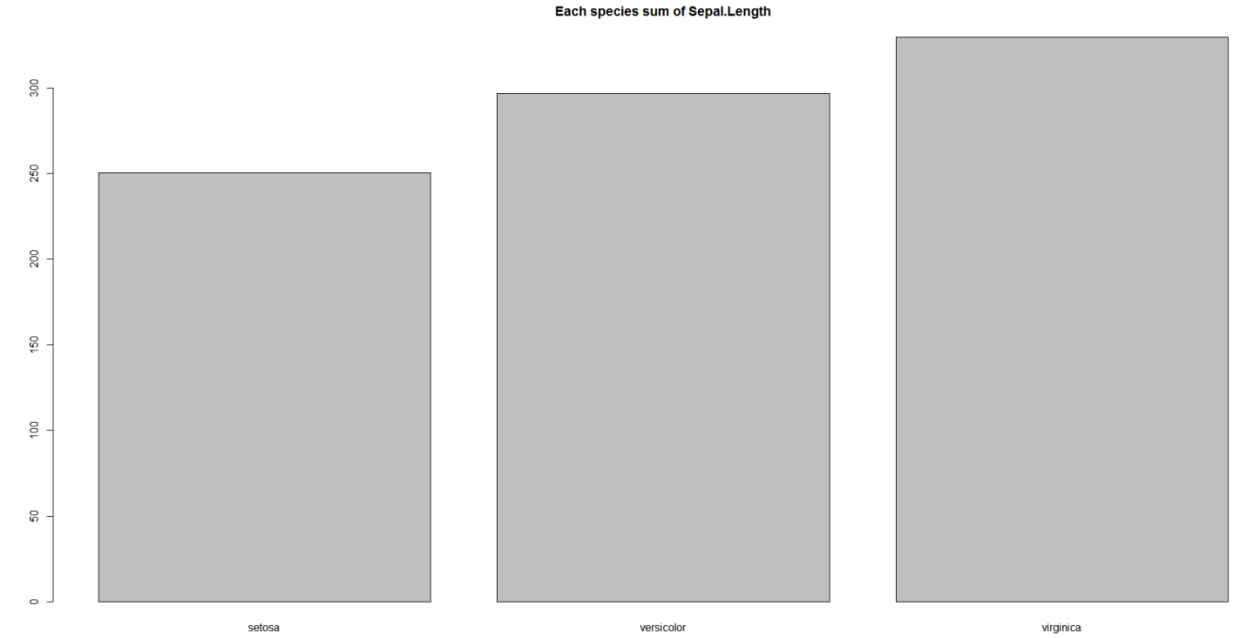
- ***Visualizing Aggregate Values with Bar plots and Pie charts***
  - sum(Sepal.Length[which(Species == "setosa")]) # The sum of setosa's Sepal.Length

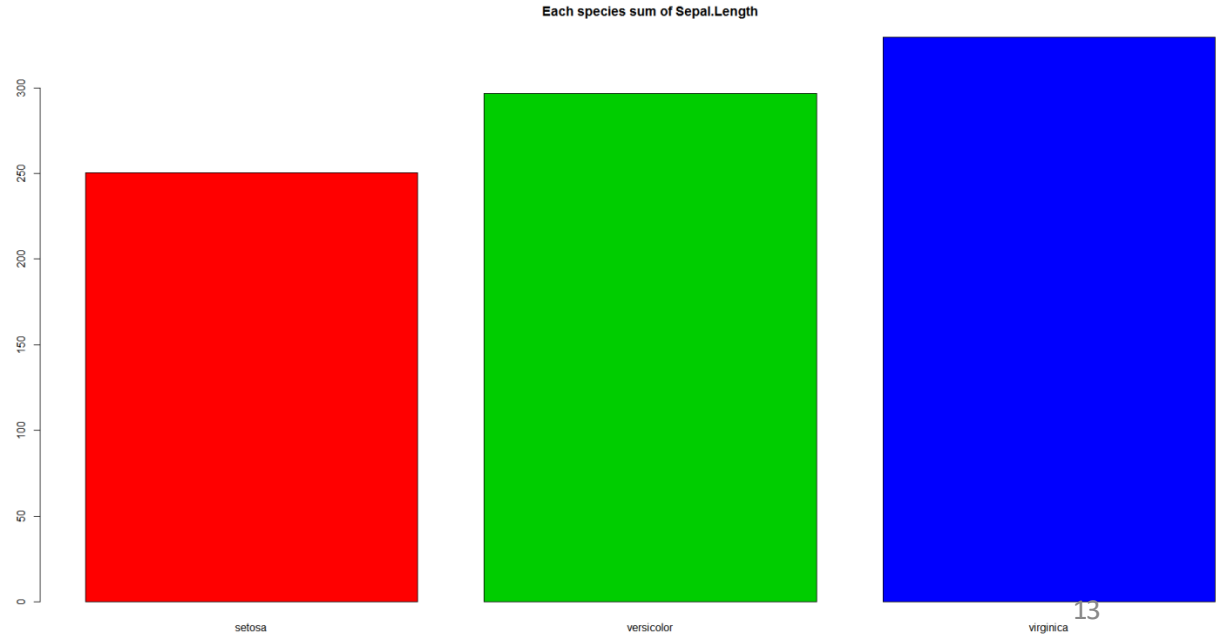
    ```
    [1] 250.3
    ```
  - by(Sepal.Length,Species,sum)#sum of each species's Sepal.Length

```
Species: setosa
[1] 250.3
------------------------------------------------------------
Species: versicolor
[1] 296.8
------------------------------------------------------------
Species: virginica
[1] 329.4
```

Visualizing Aggregate Values with Bar plots and Pie charts

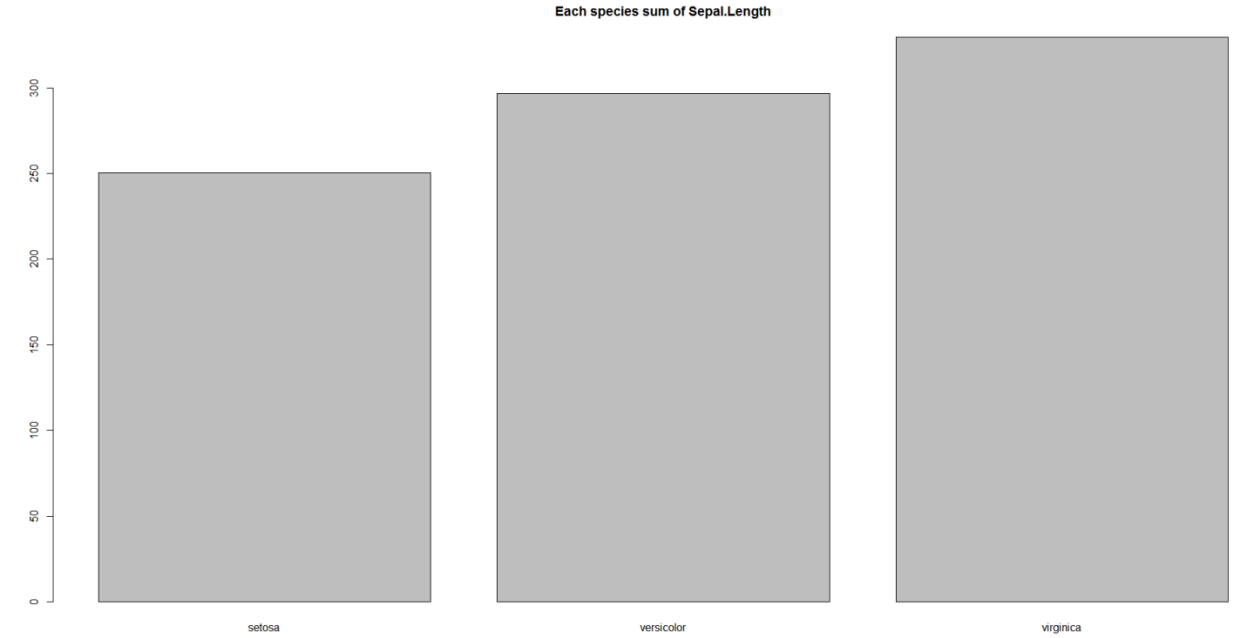
- ***Visualizing Aggregate Values with Bar plots***
  - barplot(by(Sepal.Length,Species,sum),main = "each species sum of Sepal.Length")
  - barplot(by(Sepal.Length,list(Species),sum),col=2:4,main = "Each species sum of Sepal.Length")

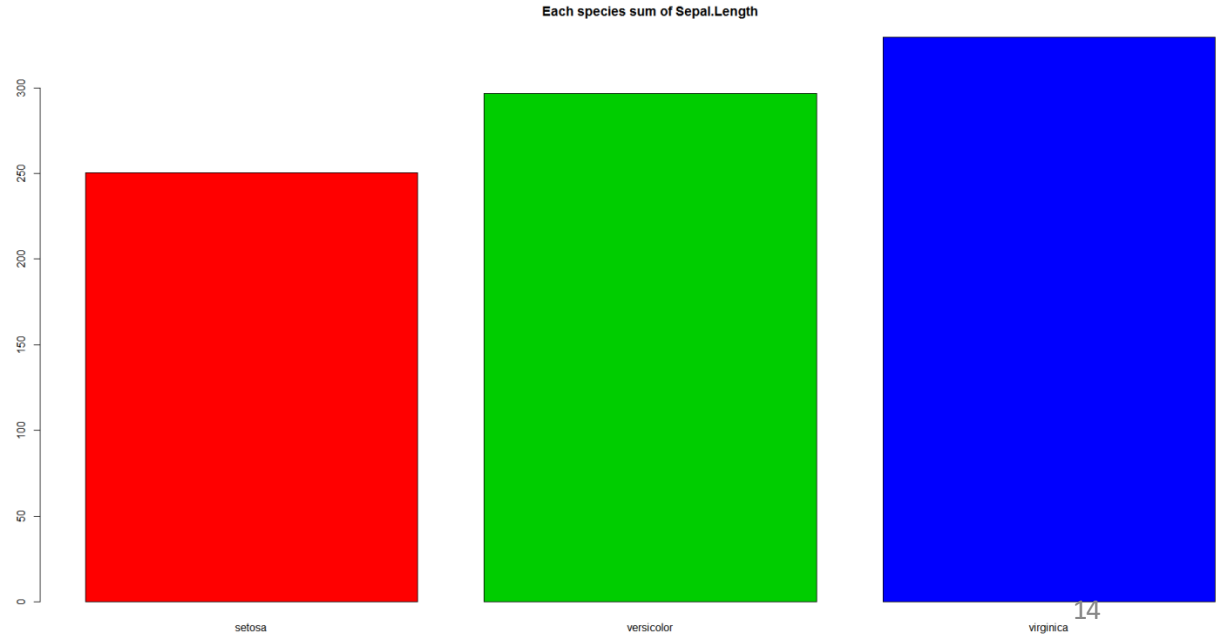




Visualizing Aggregate Values with Bar plots and Pie charts

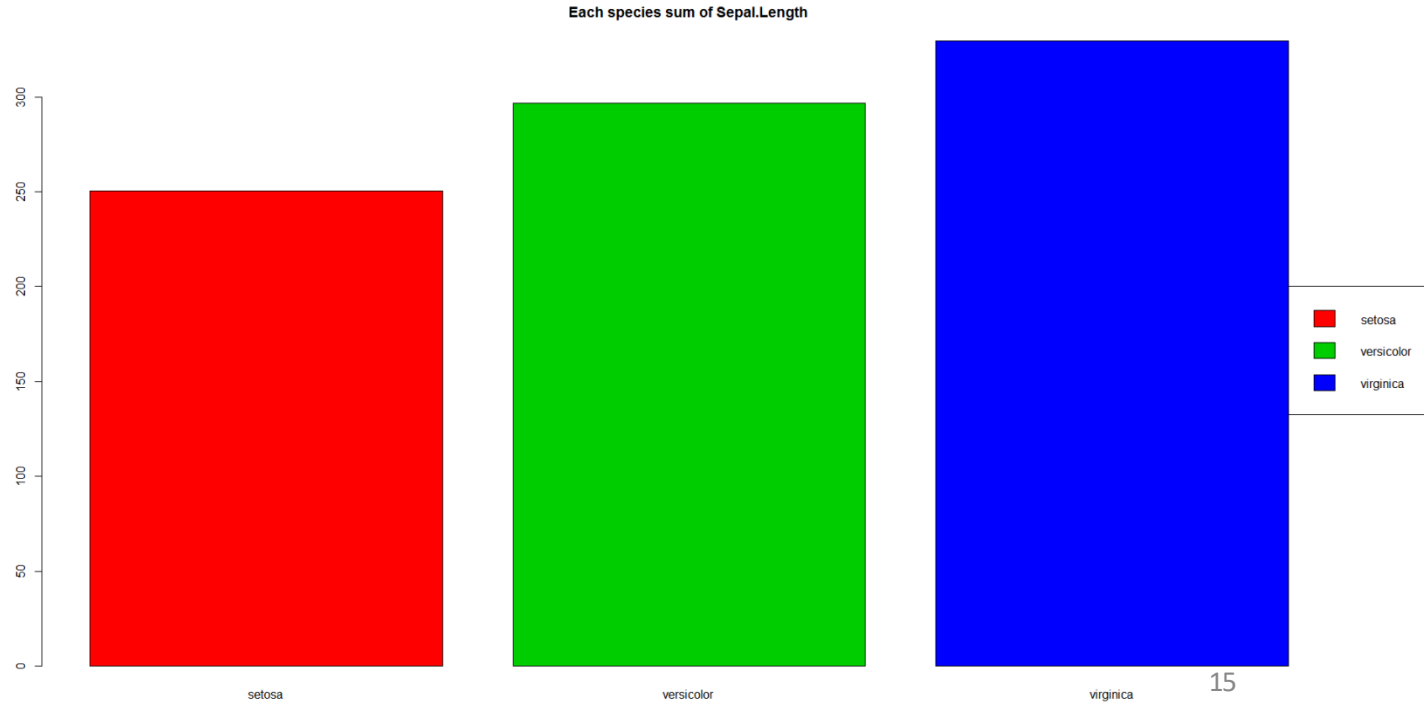
- ***Visualizing Aggregate Values with Bar plots***
  - barplot(by(Sepal.Length,Species,sum),main = "each species sum of Sepal.Length")
  - barplot(by(Sepal.Length,list(Species),sum),col=2:4,main = "Each species sum of Sepal.Length")





Visualizing Aggregate Values with Bar plots and Pie charts

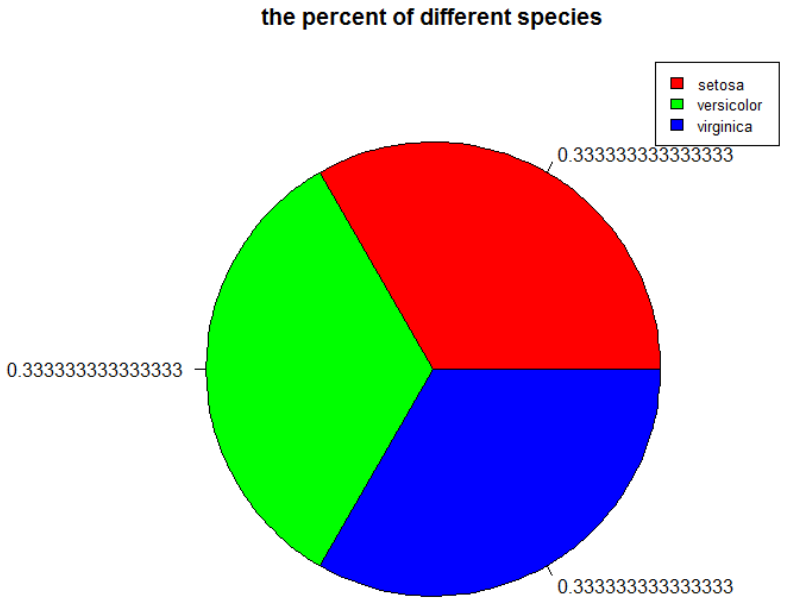
- ***Visualizing Aggregate Values with Bar plots***
  - par(xpd=T, mar=par()$mar + c(0,0,0,4))
  - barplot(by(Sepal.Length,list(Species),sum),col=2:4,beside=T,main = "Each species sum of Sepal.Length")
  - legend(3.7,200,c("setosa","versicolor","virginica"),fill=2:4)



Visualizing Aggregate Values with Bar plots and Pie charts

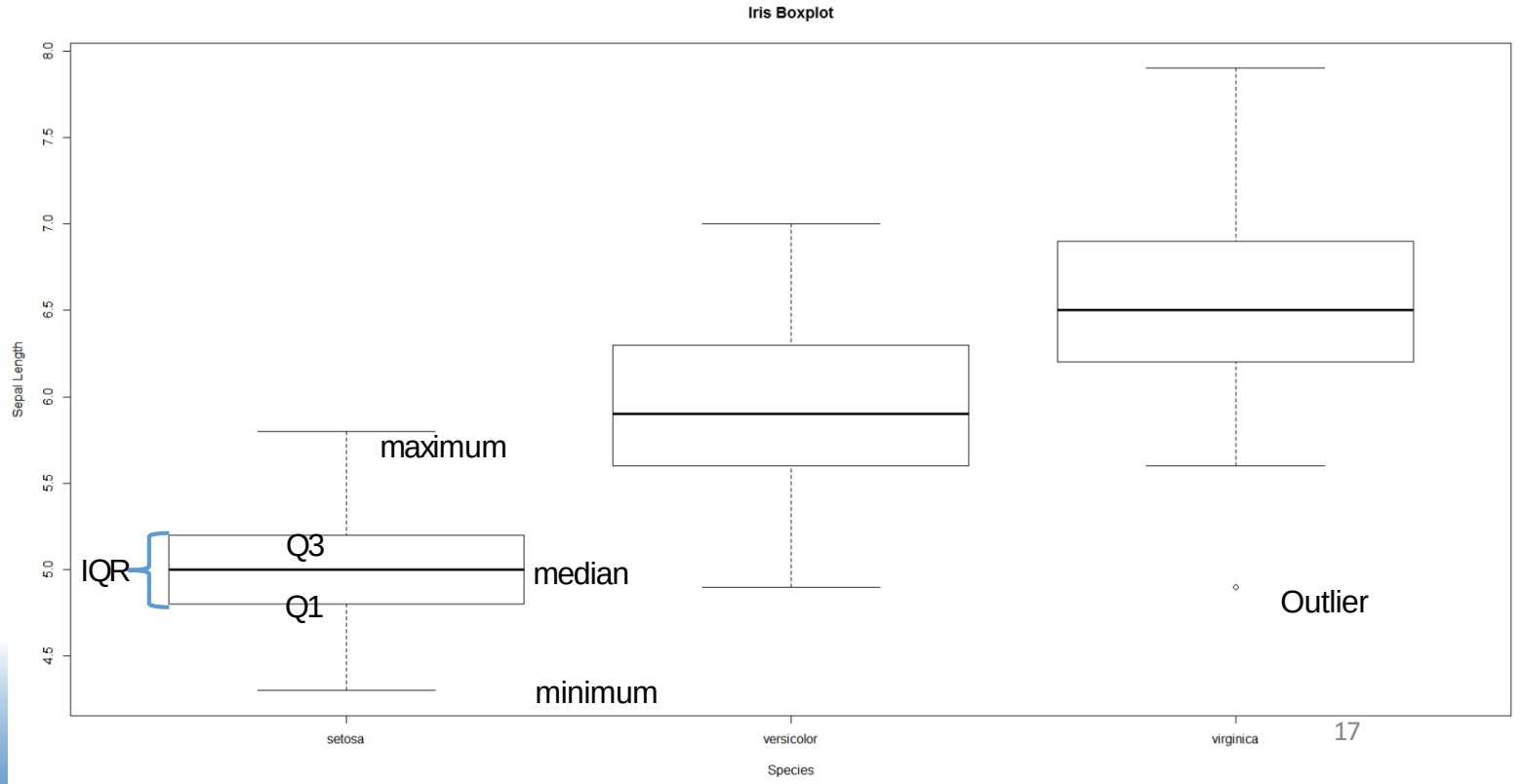
- ***Visualizing Aggregate Values with Pie charts***

- label_sp = table(Species)/sum(table(Species))
- pie(table(Species),label = label_sp,col = rainbow(length(table(Species))),main = "the percent of different species")
- legend("topright", c("setosa","versicolor","virginica"), cex = 0.8,
- fill = rainbow(length(table(Species))))



Other visualization

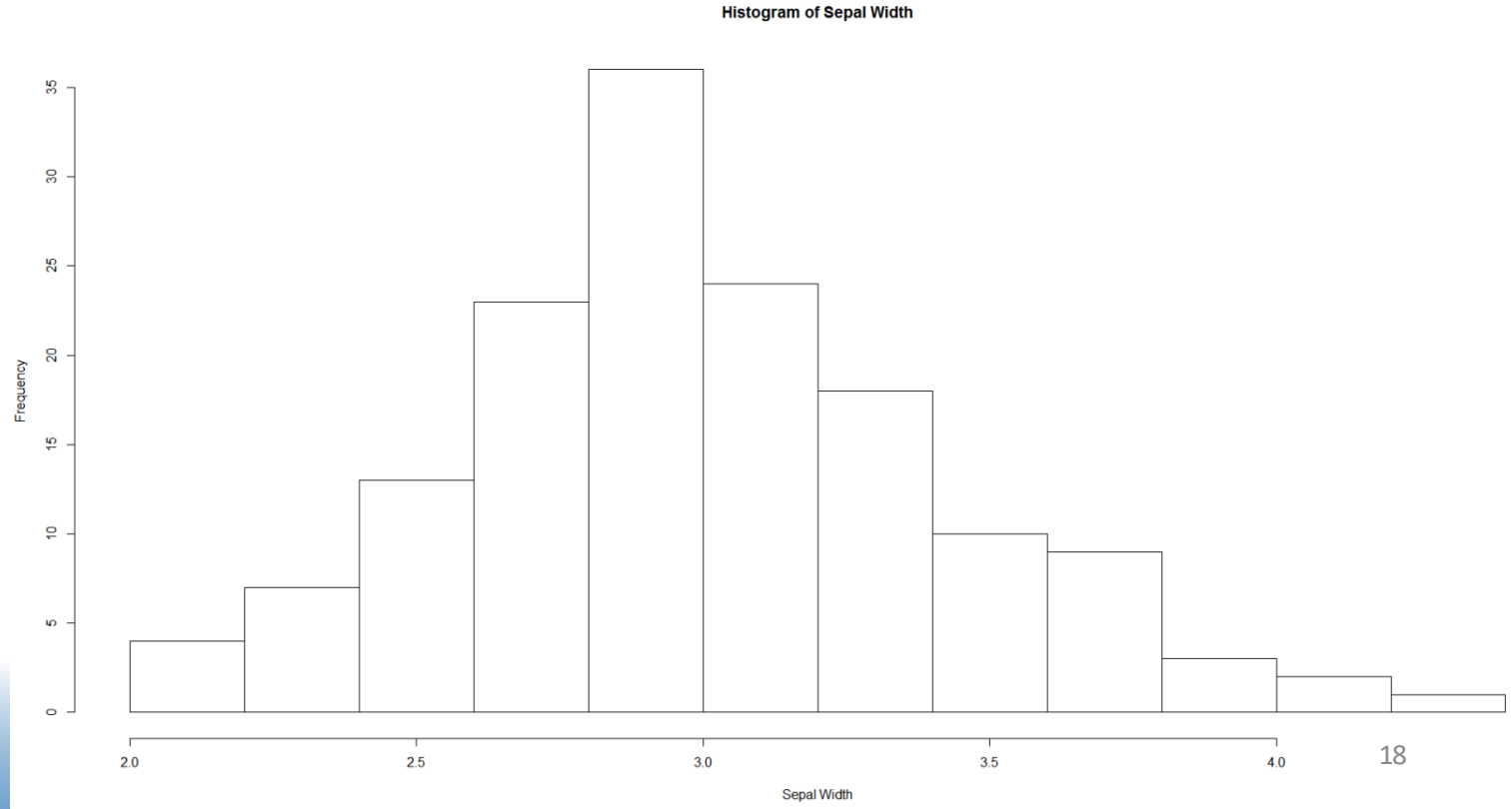
- ***Box Plot***
  - Basic setup
    - The box plot can show the minimum, Q1, median(Q2), Q3, maximum, and IQR(Q3 – Q1)
    - The value more than Q3+1.5*IQR or Q1-1.5*IQR can be view asoutlier
  - Code:
    - boxplot(Sepal.Length~Species,data=iris, xlab="Species", ylab="Sepal Length", main="Iris Boxplot")



Other visualization

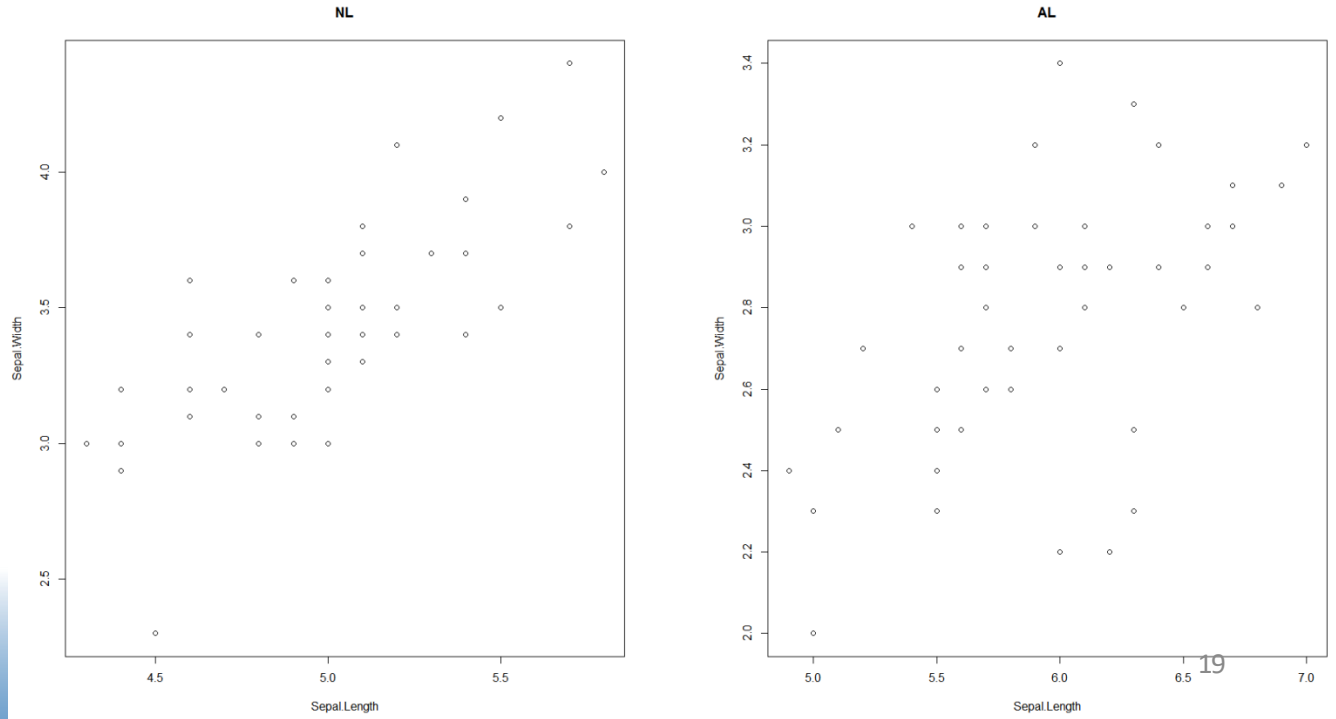
- ***Histogram***
  - Basic setup

    The histogram need to decide the numbers of bins (breaks)
  - code
  - hist(iris$Sepal.Width, freq=NULL, density=NULL, breaks=12, xlab="Sepal Width", ylab="Frequency", main="Histogram of Sepal Width")



Common Plotting Tasks

- **Multiple Plots**
  - par(mfrow=c(1,2))
  - s1 = which(Species=="setosa");s2 = which(Species=="versicolor");
  - plot(Sepal.Length[s1],Sepal.Width[s1],main="NL",xlab="Sepal.Length",ylab="Sepal.Width")
  - plot(Sepal.Length[s2],Sepal.Width[s2],main="AL",xlab="Sepal.Length",ylab="Sepal.Width")



Common Plotting Tasks

- **Saving Plots to Files**
  - getwd()#workspace

```
> getwd()
[1] "C:/Users/USER/Documents"
```

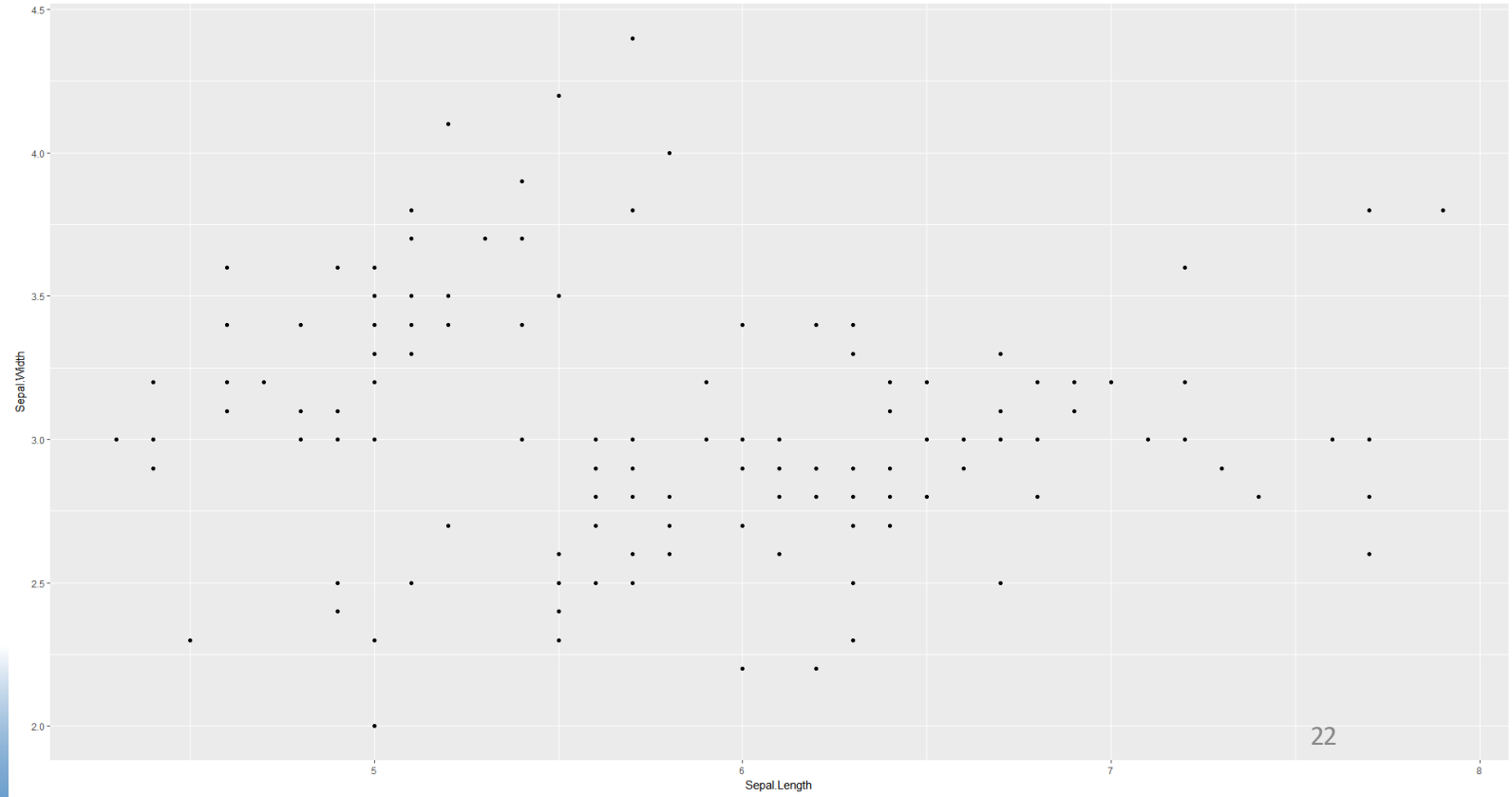
  - pdf("plot.pdf",width=6,height=4,paper='special')
  - par(mfrow=c(1,2))
  - s1 = which(Species=="setosa");s2 = which(Species=="versicolor");
  - plot(Sepal.Length[s1],Sepal.Width[s1],main="NL",xlab="Sepal.Length", ylab="Sepal.Width")
  - plot(Sepal.Length[s2],Sepal.Width[s2],main="AL",xlab="Sepal.Length", ylab="Sepal.Width")
  - dev.off() #the file will be saved in the work space

Layered Visualizations Using ggplot2

- ***Creating Plots Using qplot()***
- ***ggplot(): Specifying the Grammar of the Visualization***
- ***Themes***

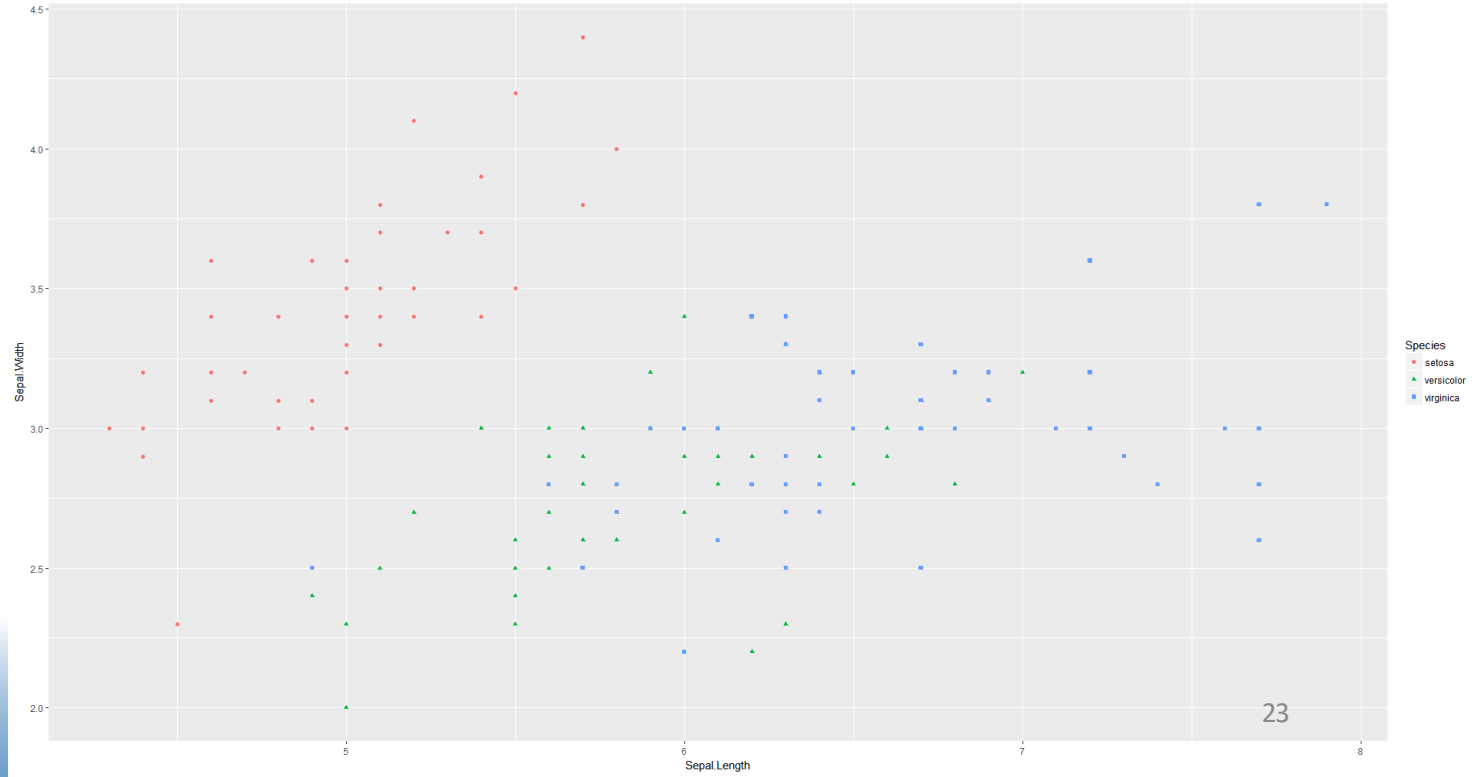
Layered Visualizations Using ggplot2

- ***Creating Plots Using qplot()***
  - install.packages("ggplot2")
  - library(ggplot2)
  - qplot(Sepal.Length,Sepal.Width)



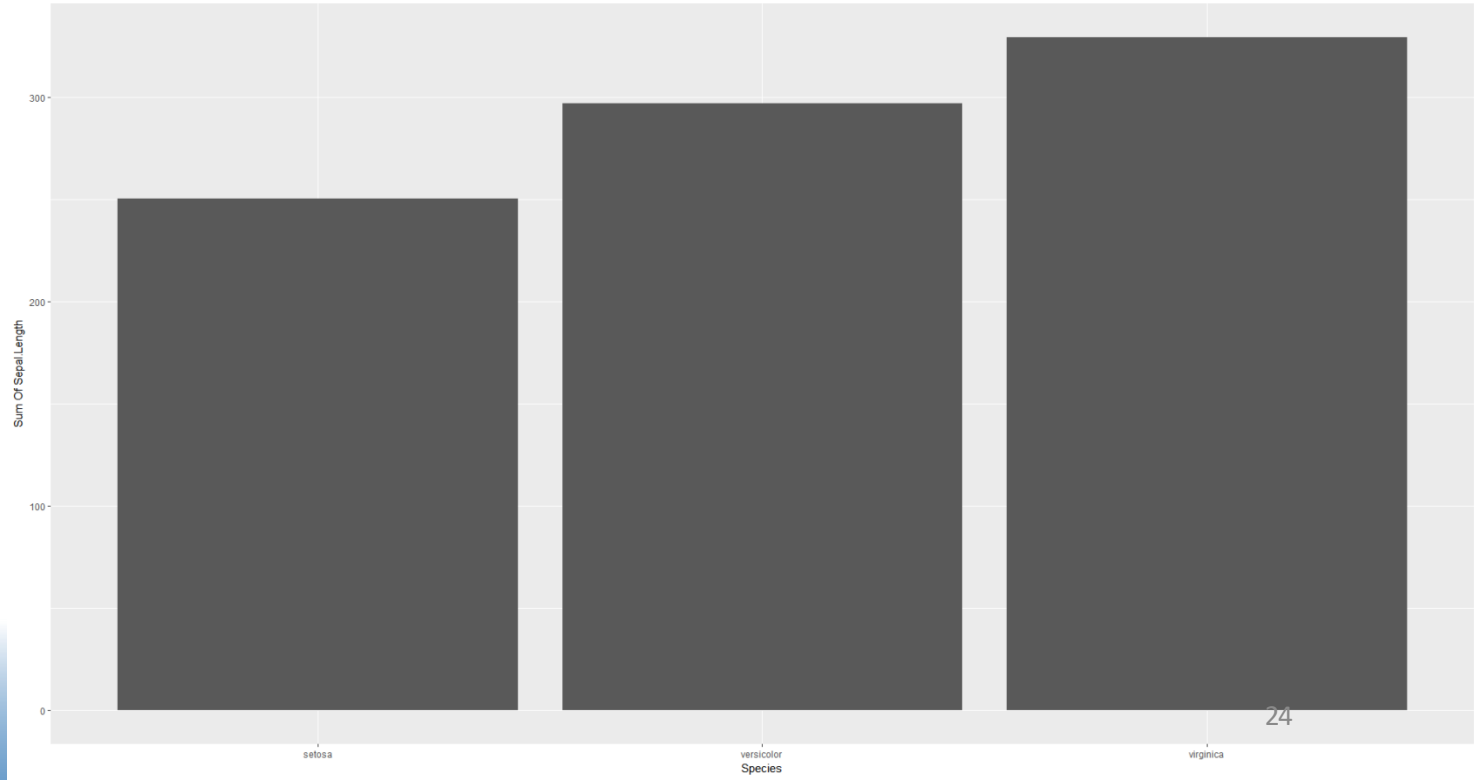
Layered Visualizations Using ggplot2

- ***Creating Plots Using qplot()***
  - install.packages("ggplot2")
  - library(ggplot2)
  - qplot(Sepal.Length,Sepal.Width,col=Species,shape=Species)



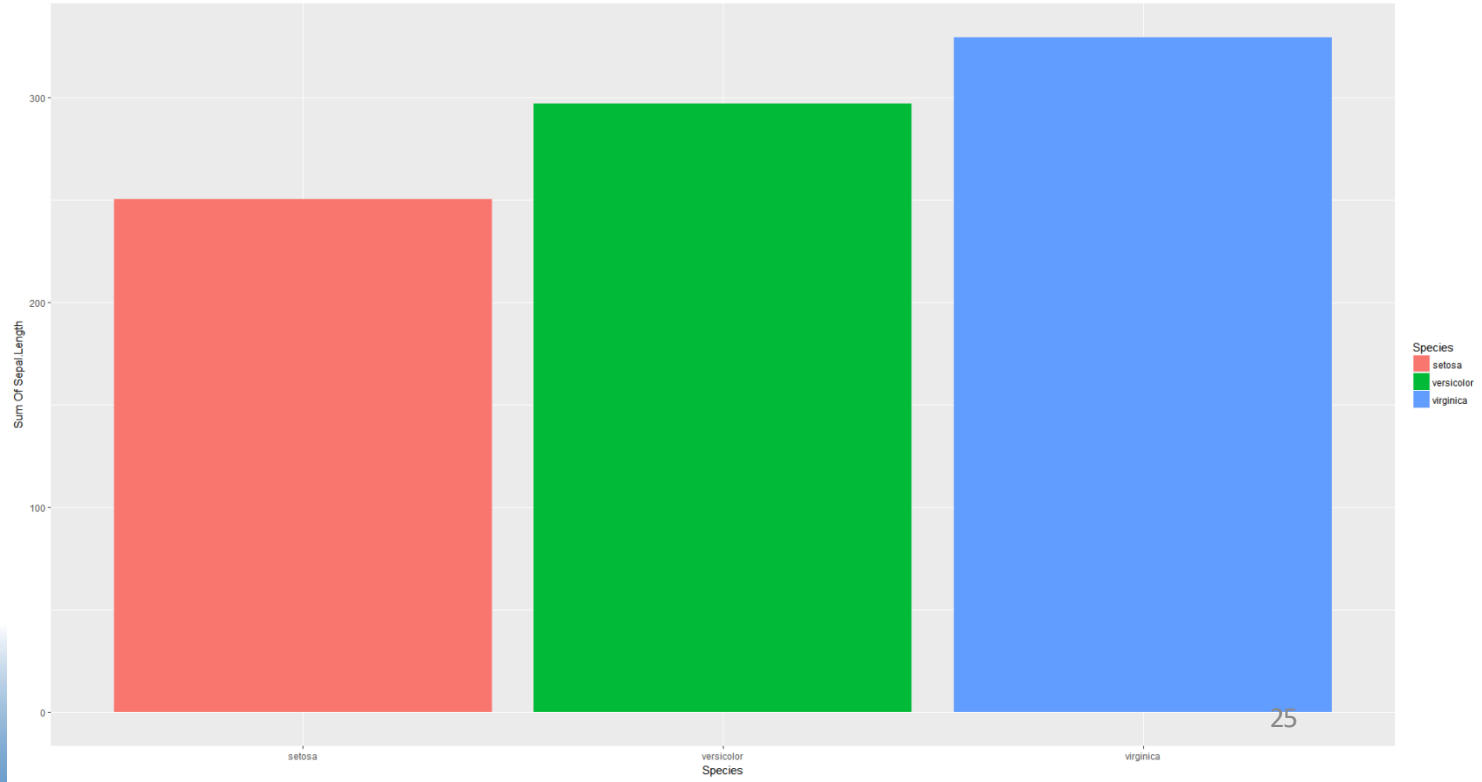
Layered Visualizations Using ggplot2

- ***Creating Plots Using qplot()***
  - install.packages("ggplot2")
  - library(ggplot2)
  - qplot(Species,weight=Sepal.Length,ylab="Sum Of Sepal.Length")



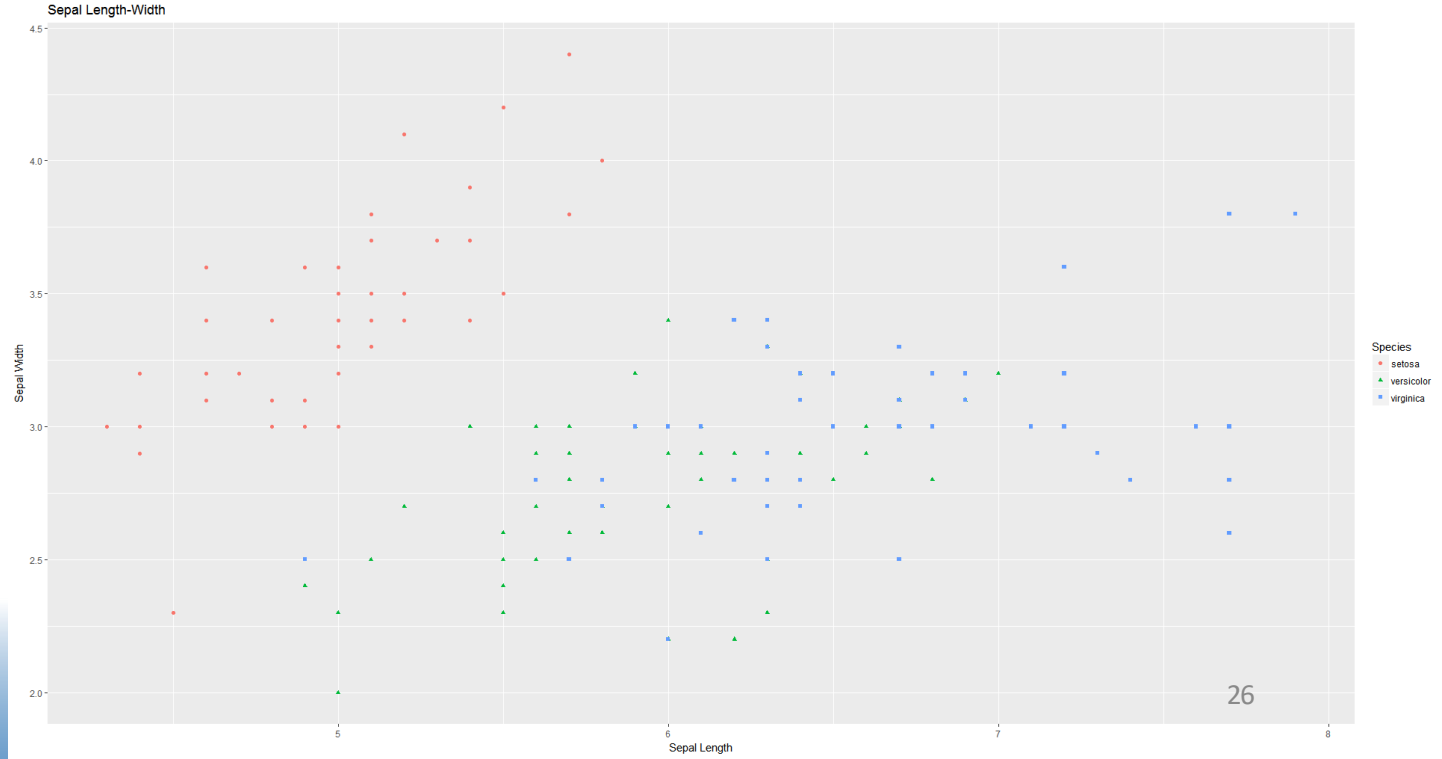
Layered Visualizations Usingggplot2

- ***Creating Plots Using qplot()***
  - install.packages("ggplot2")
  - library(ggplot2)
  - qplot(Species,weight=Sepal.Length,ylab="Sum Of Sepal.Length",fill=Species)



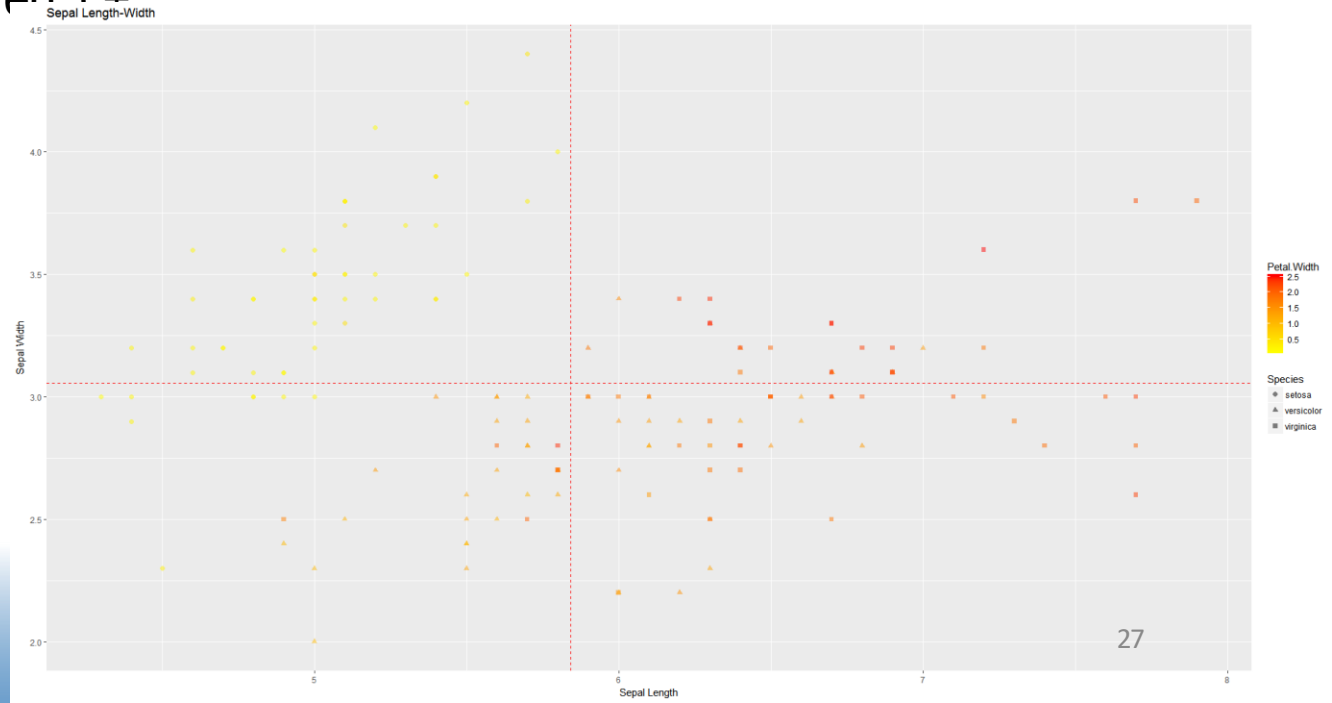
Layered Visualizations Using ggplot2

- **Layered Visualizations Using ggplot2**
  - scatter <- ggplot(data=iris, aes(x = Sepal.Length, y = Sepal.Width))
  - scatter + geom_point(aes(color=Species, shape=Species)) + xlab("Sepal Length") + ylab("Sepal Width")
    - + ggtitle("Sepal Length-Width")



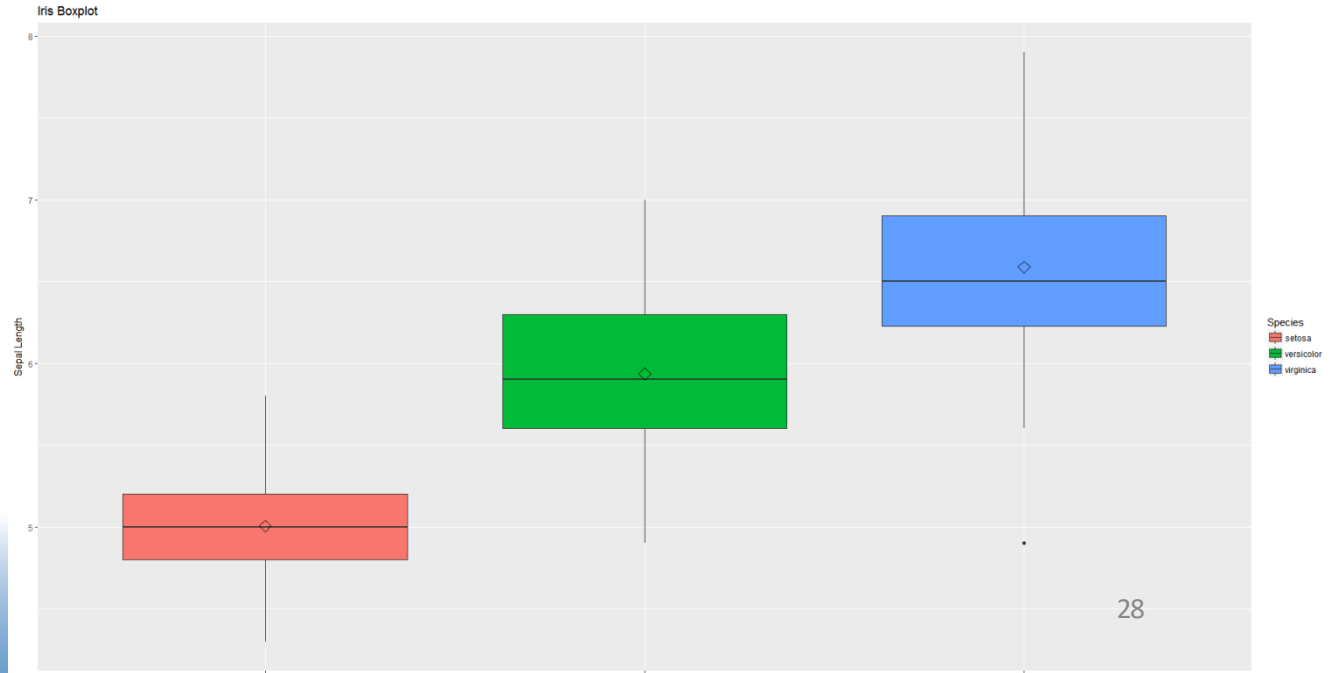
Layered Visualizations Using ggplot2

- **Layered Visualizations Using ggplot2**
  - scatter + geom_point(aes(color = Petal.Width, shape = Species), size = 2, alpha = I(1/2)) +
  - geom_vline(aes(xintercept = mean(Sepal.Length)), color = "red", linetype = "dashed") +
  - geom_hline(aes(yintercept = mean(Sepal.Width)), color = "red", linetype = "dashed") +
  - scale_color_gradient(low = "yellow", high = "red") +
  - xlab("Sepal Length") + ylab("Sepal Width") +
  - ggtitle("Sepal Length-Width")



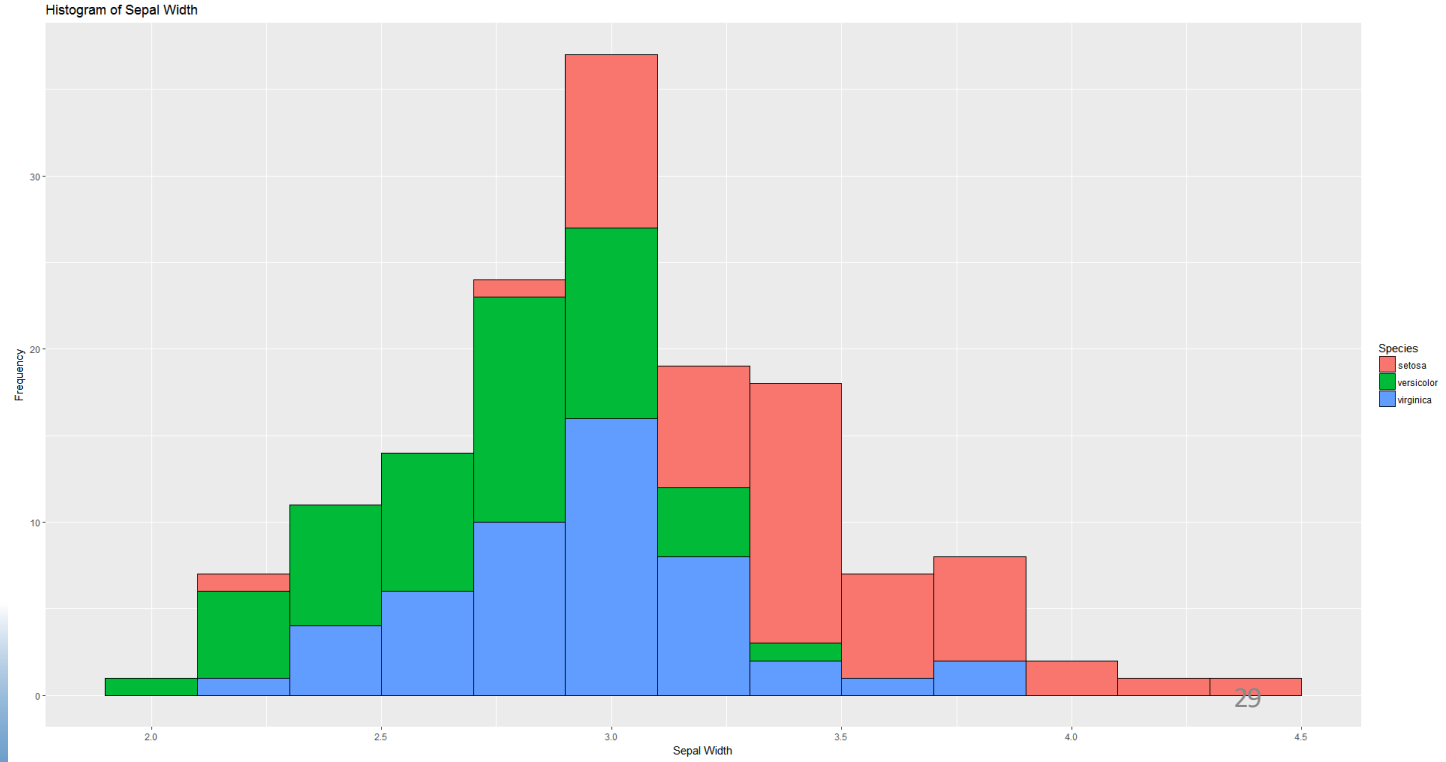
Other visualization Using ggplot2

- ***Box Plot***
  - box <- ggplot(data=iris, aes(x=Species, y=Sepal.Length))
  - box + geom_boxplot(aes(fill=Species)) + ylab("Sepal Length") + ggtitle("Iris Boxplot") +
  - stat_summary(fun.y=mean, geom="point", shape=5, size=4)



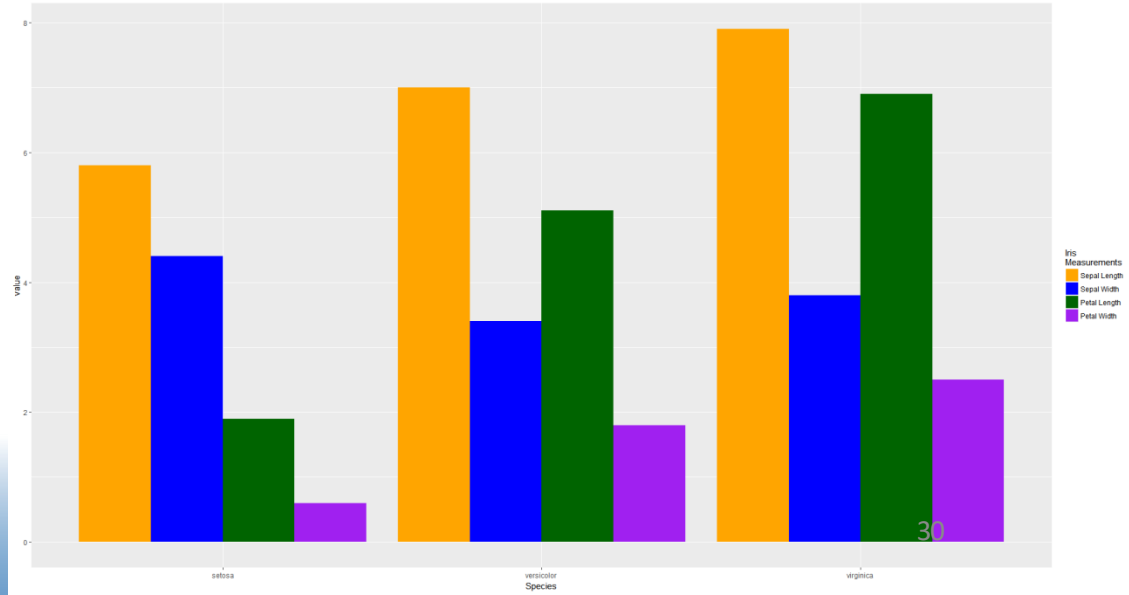
Other visualization Using ggplot2

- ***Histogram***
  - histogram <- ggplot(data=iris, aes(x=Sepal.Width))
  - histogram + geom_histogram(binwidth=0.2, color="black", aes(fill=Species)) + xlab("Sepal Width") + ylab("Frequency") + ggtitle("Histogram of SepalWidth")



Other visualization Using ggplot2

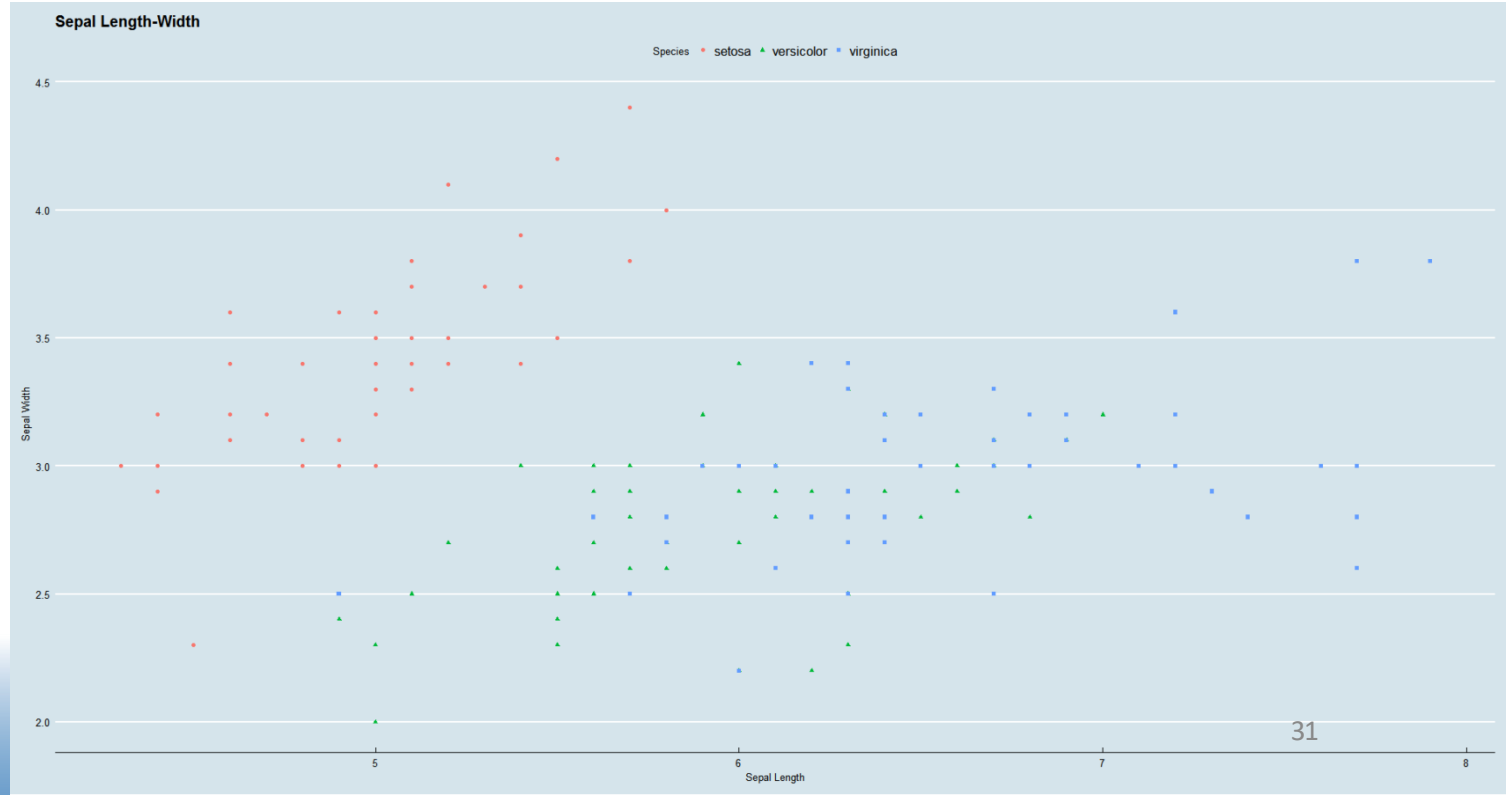
- ***Histogram***
  - library(reshape2)
  - iris2 <- melt(iris, id.vars="Species")
  - bar1 <- ggplot(data=iris2, aes(x=Species, y=value, fill=variable))
  - bar1 + geom_bar(stat="identity", position="dodge") + scale_fill_manual(values=c("orange", "blue", "darkgreen", "purple"), name="Iris\nMeasurements", breaks=c("Sepal.Length", "Sepal.Width", "Petal.Length", "Petal.Width"),labels=c("Sepal Length", "Sepal Width", "Petal Length", "Petal Width"))



Other visualization Using ggplot2

- ***Theme***

- install.packages("ggthemes")
- library(ggthemes)
- scatter + geom_point(aes(color=Species, shape=Species)) +xlab("Sepal Length") + ylab("Sepal Width") + ggtitle("Sepal Length-Width")+theme_economist()



Interactive Visualizations Using Shiny

UI

```
library(shiny)
shinyUI(fluidPage(
#fluid page for dynamically adapting to screens of different resolutions.
titlePanel("Iris Dataset"),
sidebarLayout(
sidebarPanel(
#implementing radio buttons
radioButtons("p", "Select column of iris dataset:",
list("Sepal.Length"='a', "Sepal.Width"='b', "Petal.Length"='c', "Petal.Width"='d')),
#slider input for bins of histogram
sliderInput("bins",
"Number of bins:",
min = 1,
                max = 50,
value = 30)
# Show a plot of the generated distribution
          ),
mainPanel(
plotOutput("distPlot")
)
        )
      ))
```

Interactive Visualizations Using Shiny

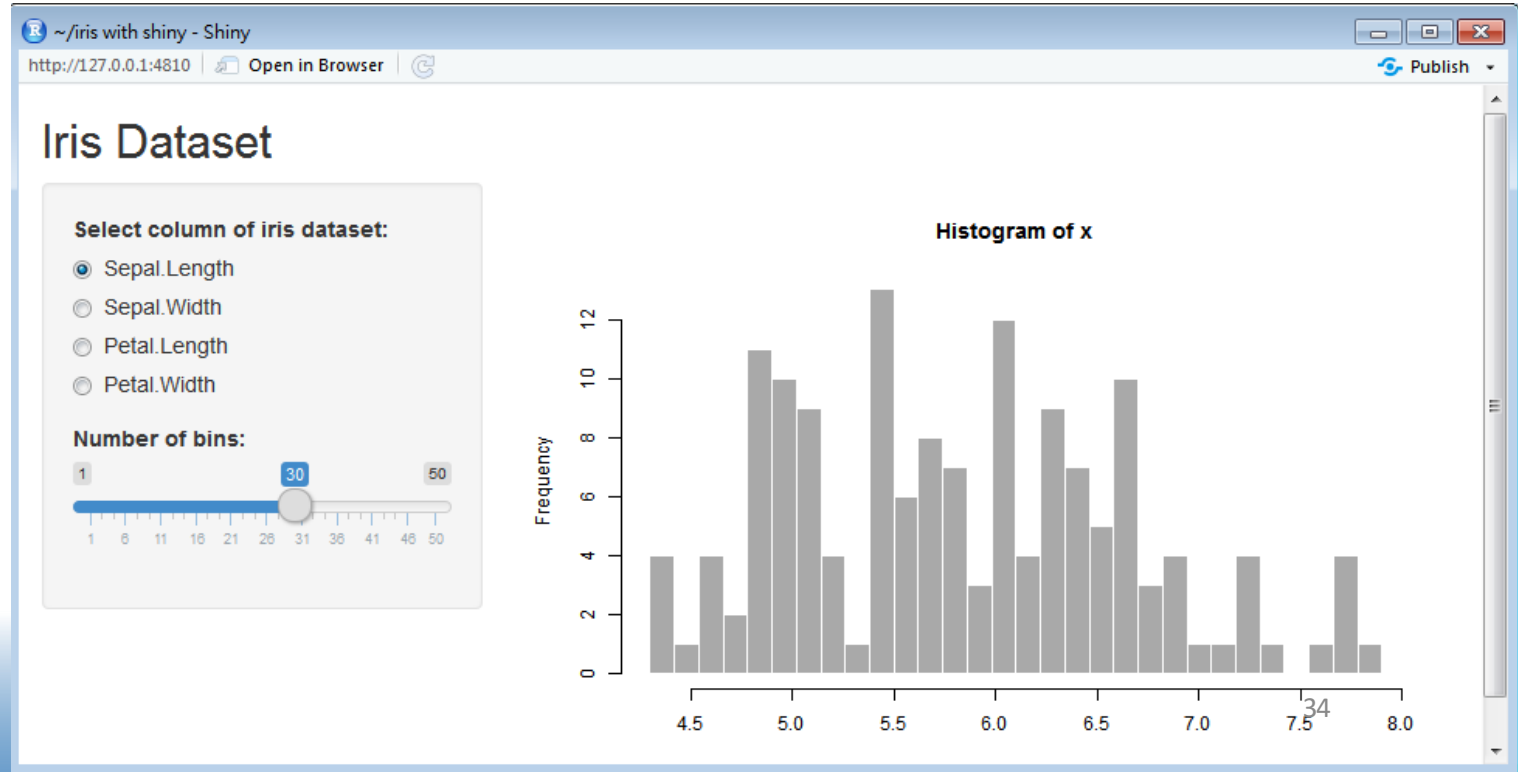
Server

```
library(shiny)
#writing server function
shinyServer(function(input, output) {
#referring output distPlot in ui.r asoutput$distPlot
output$distPlot <- renderPlot({
#referring input p in ui.r asinput$p
if(input$p=='a'){
        i<-1
       }
if(input$p=='b'){
        i<-2
       }
if(input$p=='c'){
        i<-3
}
if(input$p=='d'){
        i<-4
       }
x <- iris[, i]
#referring input bins in ui.r asinput$bins
bins <- seq(min(x), max(x), length.out = input$bins + 1)
#producing histogram as output

    hist(x, breaks = bins, col = 'darkgray', border = 'white')
})
  })
```

Interactive Visualizations Using Shiny

- Save to Rfile UI.R and Server.R into same folder
- Execute the command
  - runApp(“folder path")
  - Ex: runApp("C:/Users/USER/Documents/iris with shiny/")



3D dynamic plots with iris

- install.packages(c("rgl", "car"))
- library(car)
- attach(iris)
- scatter3d(x = iris$Sepal.Length,
- y = iris$Sepal.Width,
- z = iris$Petal.Length)
- scatter3d(x = iris$Sepal.Length,
- y = iris$Sepal.Width,
- z = iris$Petal.Length,
- groups = iris$Species)
- scatter3d(x = iris$Sepal.Length,
- y = iris$Sepal.Width,
- z = iris$Petal.Length,
- groups = iris$Species, surface=FALSE, ellipsoid = TRUE)

Homework 4

- Basic
  - Use the data you prepared to do visualized (From the course teach)
  - Boxplot, barplot, scatterplot, histogram…
- Advanced
  - Explain what you find and why you choose these visualized methods

Homework 4 (submitted to e3new.nctu.edu.tw before Oct 15, 2019)

- Use R and/or the other software to visualize the data set with missing data (NA) that you select
- Explain the results you obtain
- Discuss possible problems you plan to investigate for future studies
- Possible source of open data:

  UCI Machine Learning Repository

  (https://archive.ics.uci.edu/ml/datasets.php)

References

1. http://blog.revolutionanalytics.com/2011/03/how-the-new-york-times-uses-r-for-data-visualization.html
2. http://www.stevefenton.co.uk/Content/Pie-Charts-Are-Bad/.
3. Lewis, M. (2004). *Moneyball: The art of winning an unfair game*. NewYork: W.W. Norton &Company.
4. Teach yourself shiny. http://shiny.rstudio.com/tutorial/.
5. Tufte, E. (2001). *The visual display of quantitative information*. Cheshire: Graphics Press.
6. Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*. Use R!. NewYork: Springer.
7. Wilkinson, L. (2005). *The grammar of graphics*. NewYork:Springer.
8. https://www.analyticsvidhya.com/blog/2016/10/creating-interactive-data-visualization-using-shiny-app-in-r-with-examples/
9. https://www.mailman.columbia.edu/sites/default/files/media/fdawg_ggplot2.html