Data Visualization and Graphics in R

Xin Luo

July 21, 2017

Outline

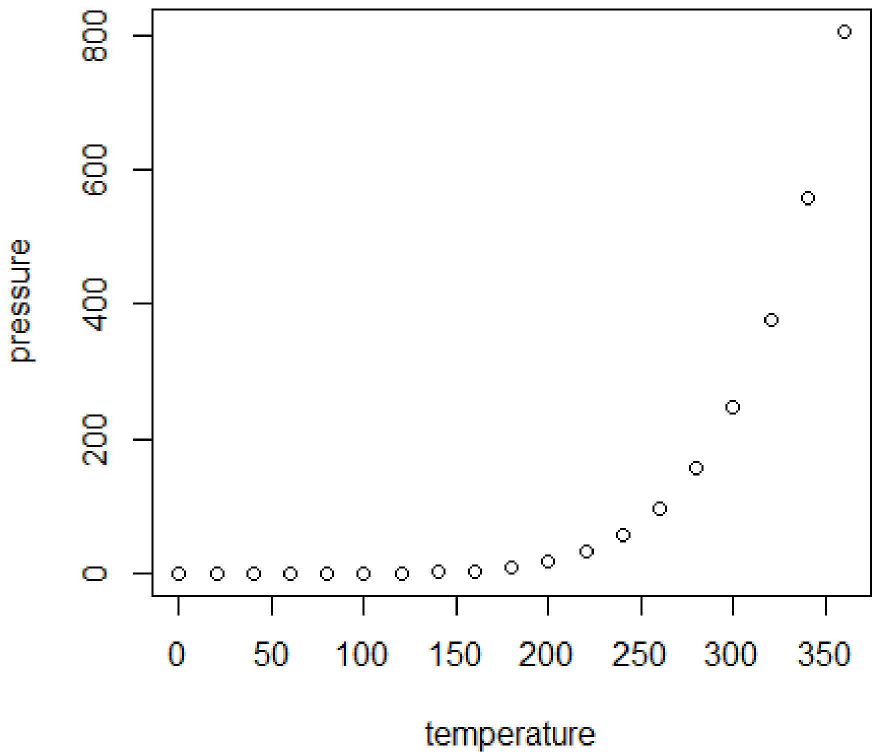
- Simple plotting using default graphics tools in R
- Plotting with graphic packages in R (ggplot2)
- Visualizing data by different types of graphs in R (scatter plot, line graph, bar graph, histogram, boxplot, pie chart, venn diagram, correlation plot, heatmap)
- Generate polished graph for publication and presentation

Why using R for plotting

1. When the large sample size exceed the capacity for excel, prism or other graphic tools
2. Fast and simple
3. Bundle with complicated statistical analysis
4. Strong graphics capabilities makes R more popular
5. Many add-on packages for various needs

Basic Plot Example

```
library ("datasets")
data ()
str (pressure)
plot (pressure)
```



Labels and Axes

Default: R uses the variable names for axes labels and computes range for axes.

Manual change by:

- axes labels: xlab, ylab
- size of labels: cex.lab
- axes range: xlim, ylim

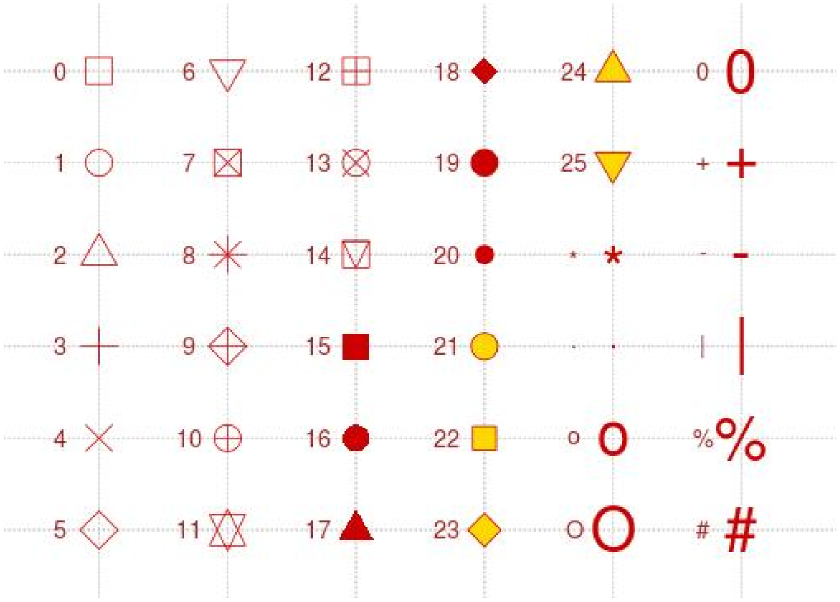
Titles

- main: sets plot title (above plot)
- sub: sets subtitle (beneath plot)

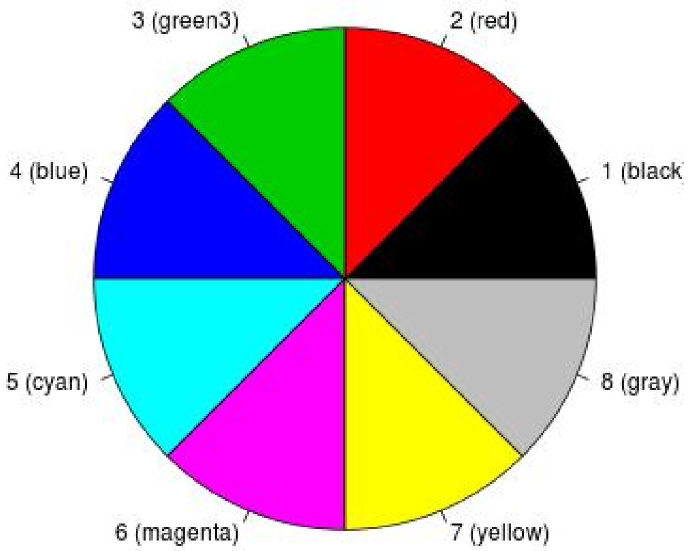
Symbols, colors and lines

- type: "p" for points, "l" for lines, "b" for both, "h" for histogram-like
- pch: point symbol
- col: color
- cex: size factor
- lty: line type
- lwd: line width

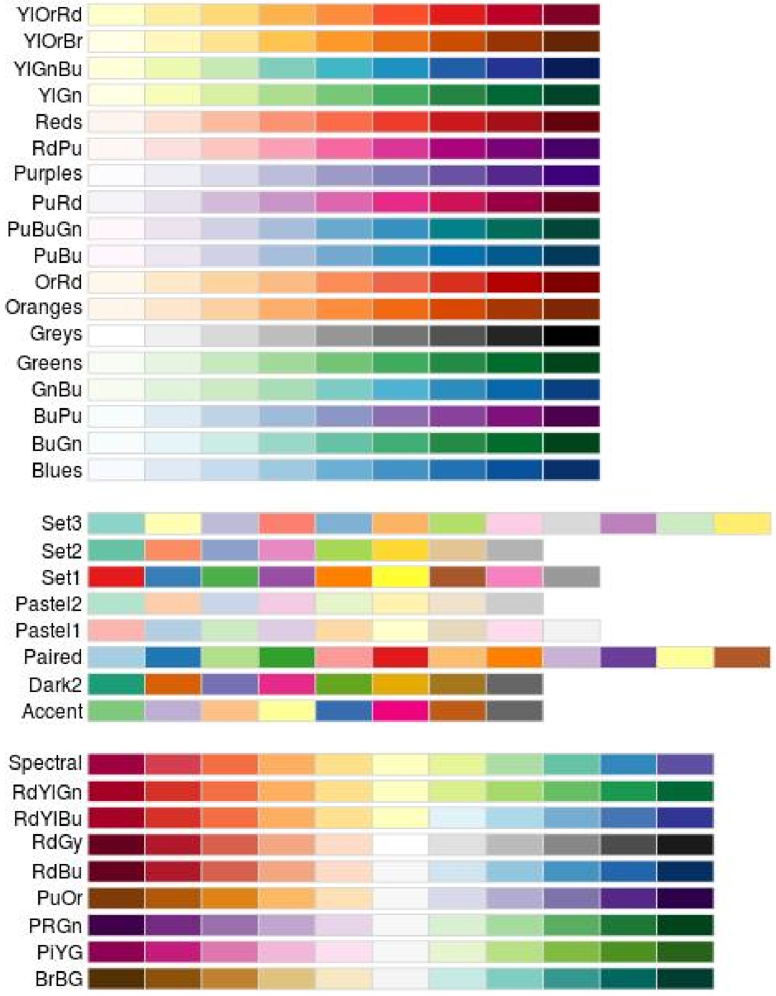
Plot symbols



The default color palette in R:



RcolorBrewer Package Palette



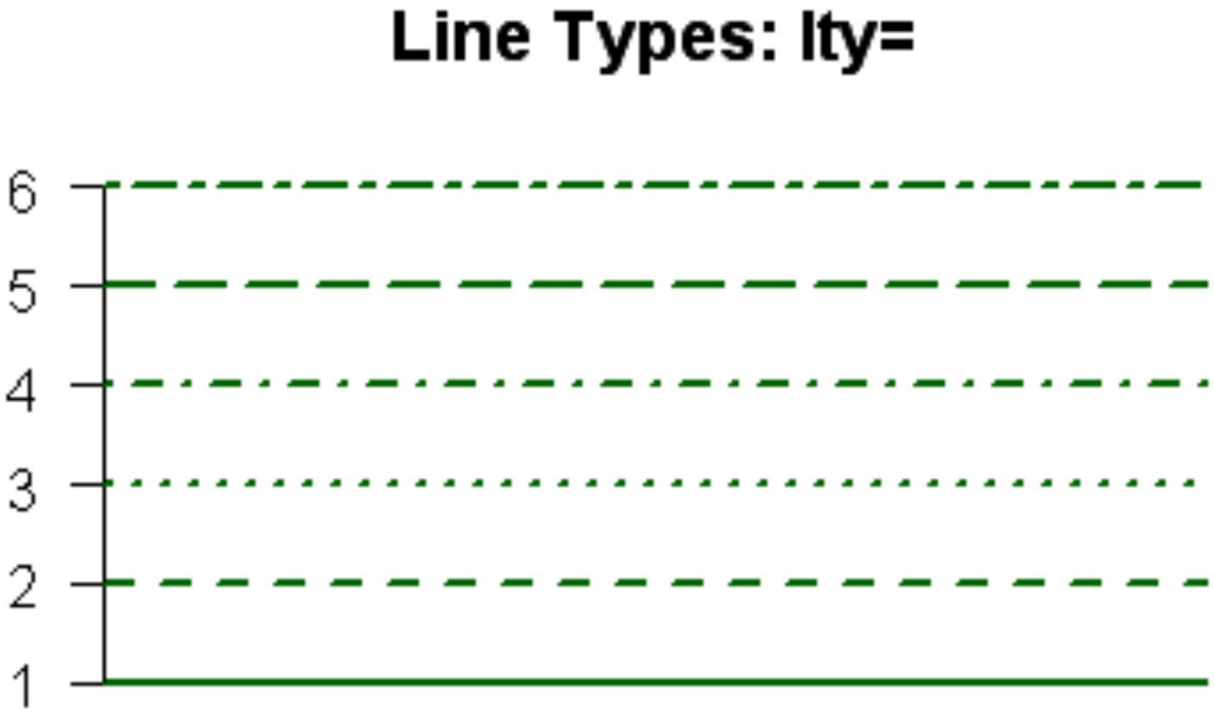
R Color Palette other options

- rainbow(n)
- heat.colors(n)
- terrain.colors(n)
- topo.colors(n)
- cm.colors(n)

Specify colors by RGB color code

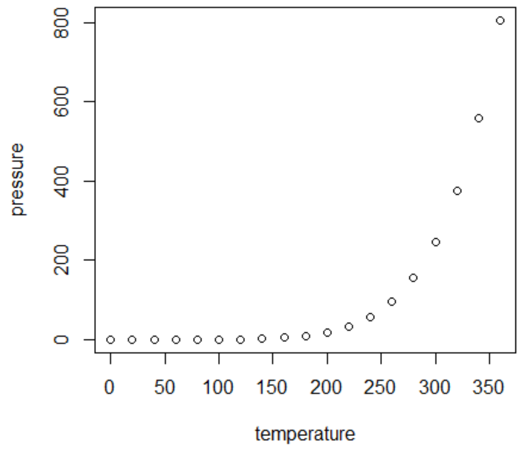
http://research.stowers.org/mcm/efg/R/Color/Chart/ColorChart.pdf

Line Types

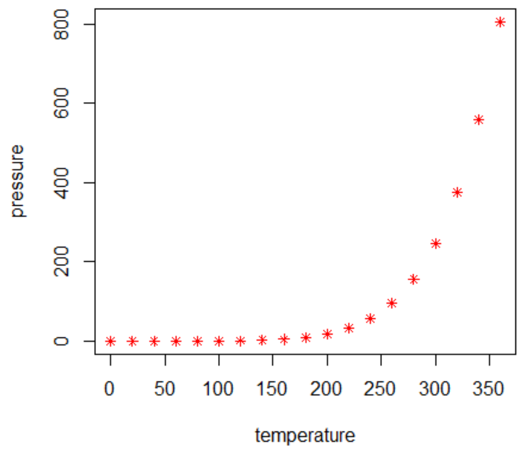


Plot example 1 plot points with formatting

```
plot (pressure, type="p")
```

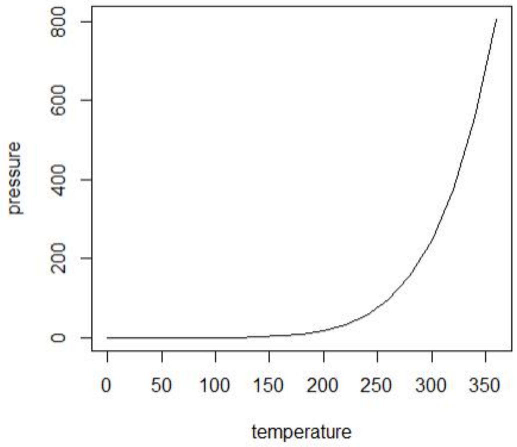


```
plot (pressure, type="p", pch = 8, cex =0.8, col="red")
```

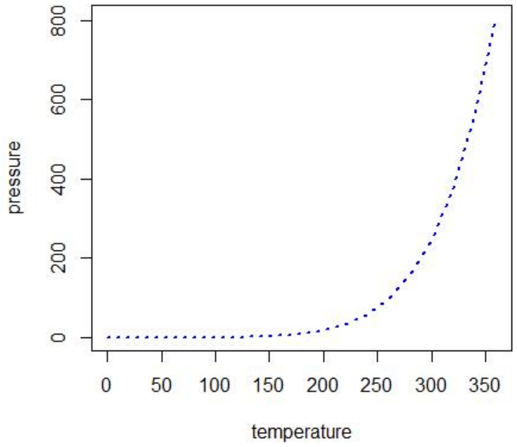


Plot example 2 line graph with formatting

```
plot (pressure, type="l")
```

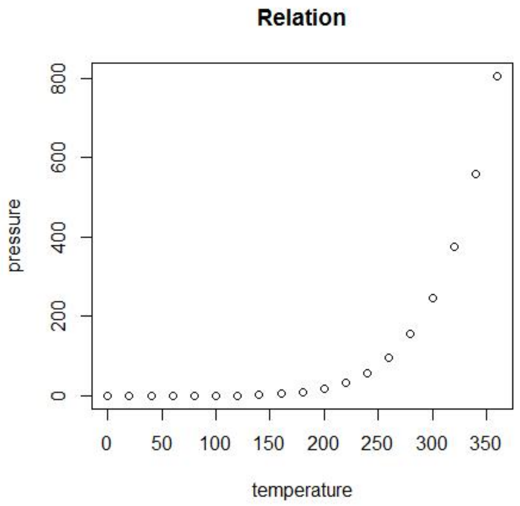


```
plot (pressure, type="l", lty=3, lwd=2, col=”blue")
```

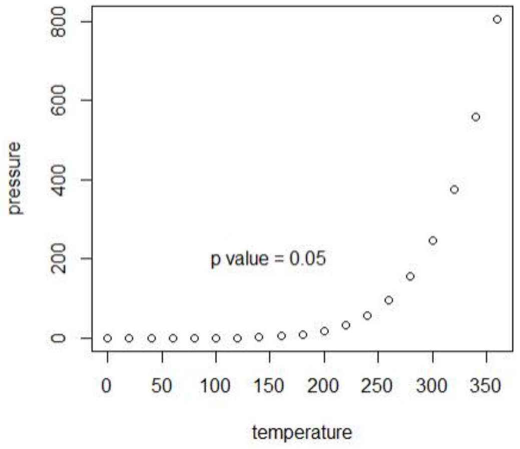


Plot example 3 add title and text

```
plot (pressure, main="Relation" )
```

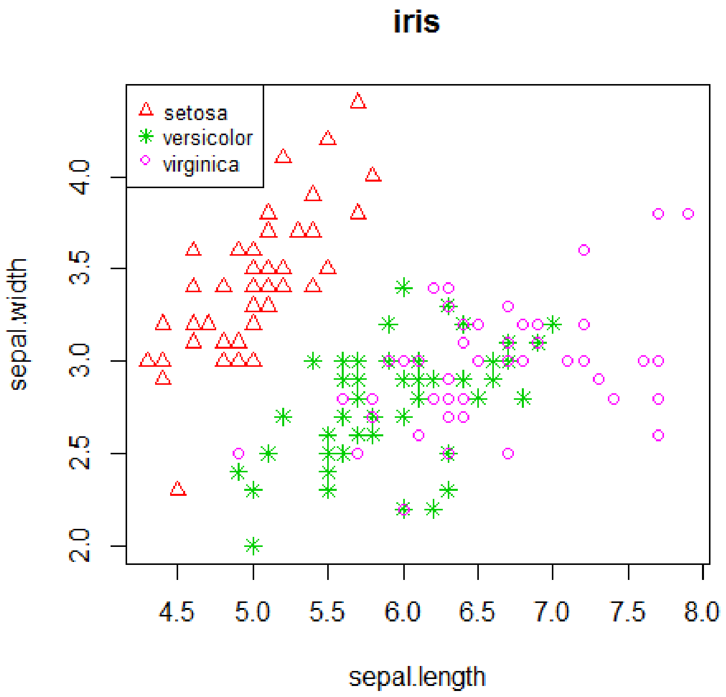


```
plot ( pressure )
text (150 ,200 , label =" p value = 0.05 ")
```



Plot for multiple group

```
data(iris) # load iris data
pch.vec <- c(2 ,8 ,21)[iris$Species]
col.vec <- c(2 ,3 ,6)[iris$Species]
plot(iris$Sepal.Length, iris$Sepal.Width, col = col.vec, pch=pch.vec, xlab="sepal.length", ylab="sepal.width",main="iris")
legend ("topleft", pch=c(2 ,8 ,21) ,col=c(2 ,3 ,6) ,legend = unique(iris$Species), cex=0.8)
```



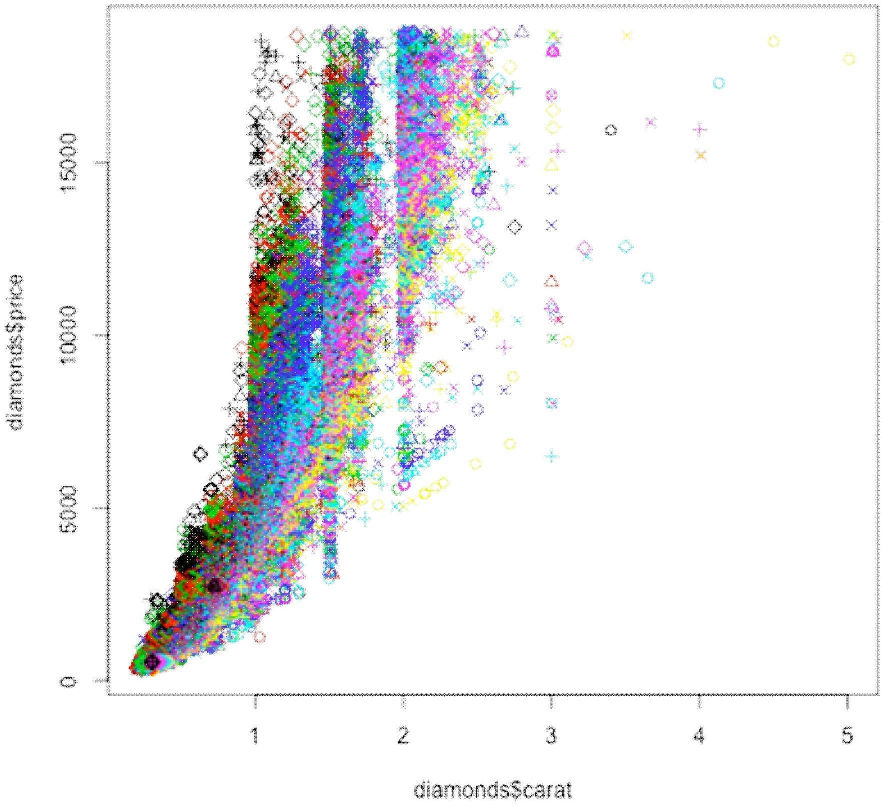
Beyond simple graphs: ggplot2

- Hadley Wickham’s ggplot2 package provides a unified interface and simple set of options.
- Once you learn how ggplot2 works for one type of plot, you can easily apply the knowledge for any other types of plots
- It provides beautiful, publication ready results.
- Easy to plot for data with multiple groups and build legend automatically

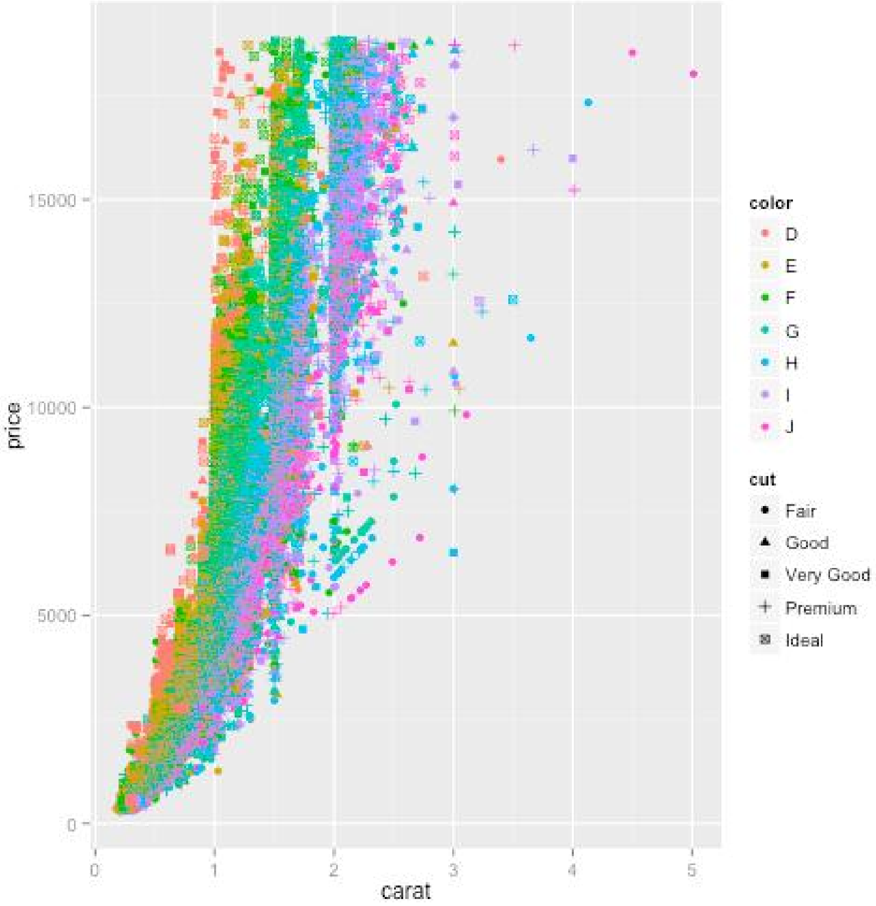
“*R Graphics Cookbook* by Winston Chang (O’Reilly). Copyright 2013 Winston Chang, 978-1-449-31695-2.”

http://www.cookbook-r.com/Graphs/

Build-in R Plotting VS ggplot2



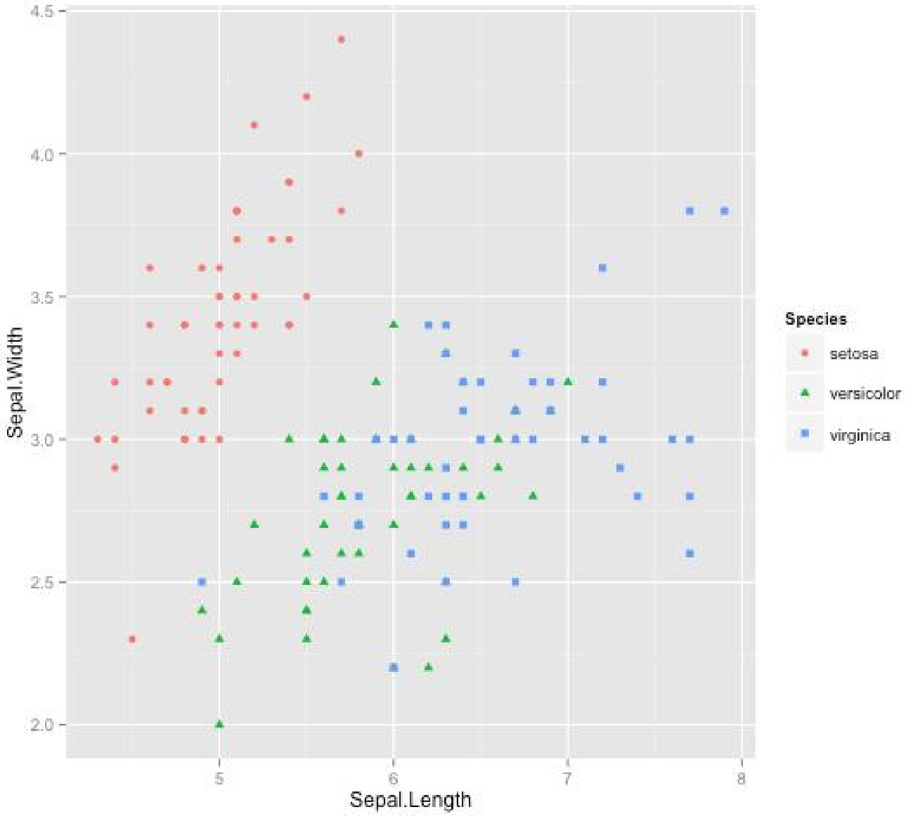
```
plot(diamonds$carat, diamonds$price, col = diamonds$color,
   pch = as.numeric(diamonds$cut))
```



```
ggplot(diamonds, aes(carat, price, col = color, shape = cut)) +
geom_point()
```

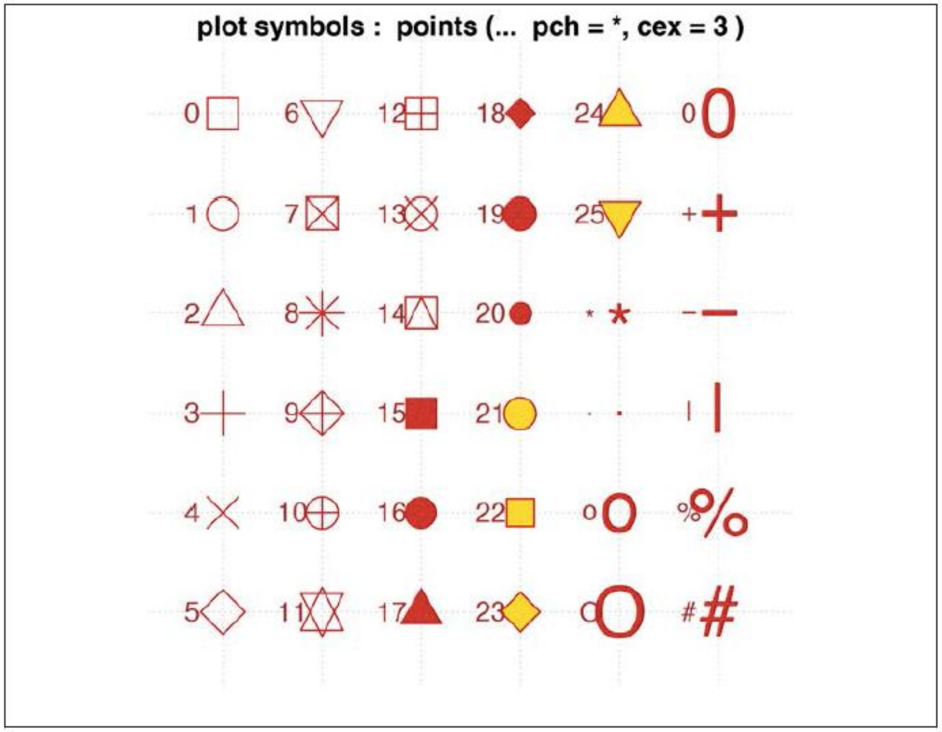
Scatter plot

display the relationship between two continuous variables

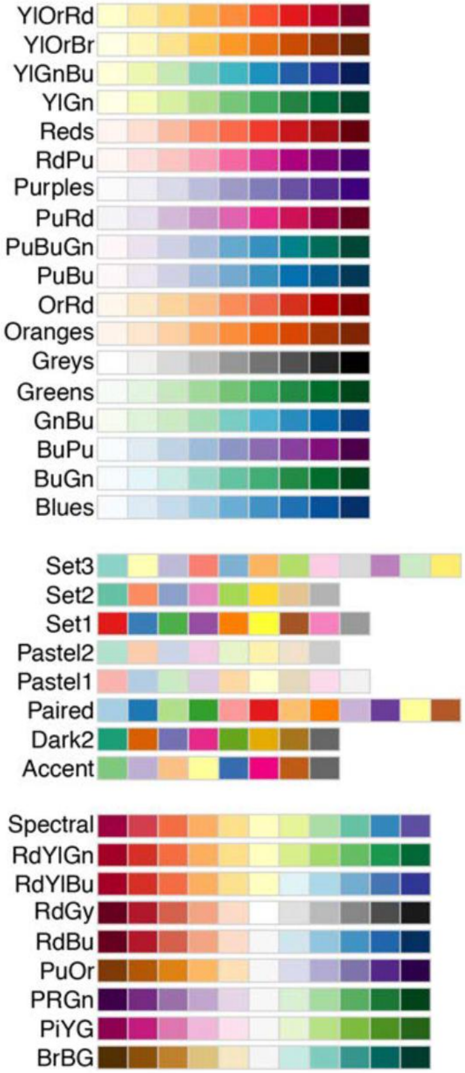


```
ggplot(iris, aes(x=Sepal.Length, y=Sepal.Width, shape=Species, colour=Species)) +
  geom_point()
```

scale_shape_manual

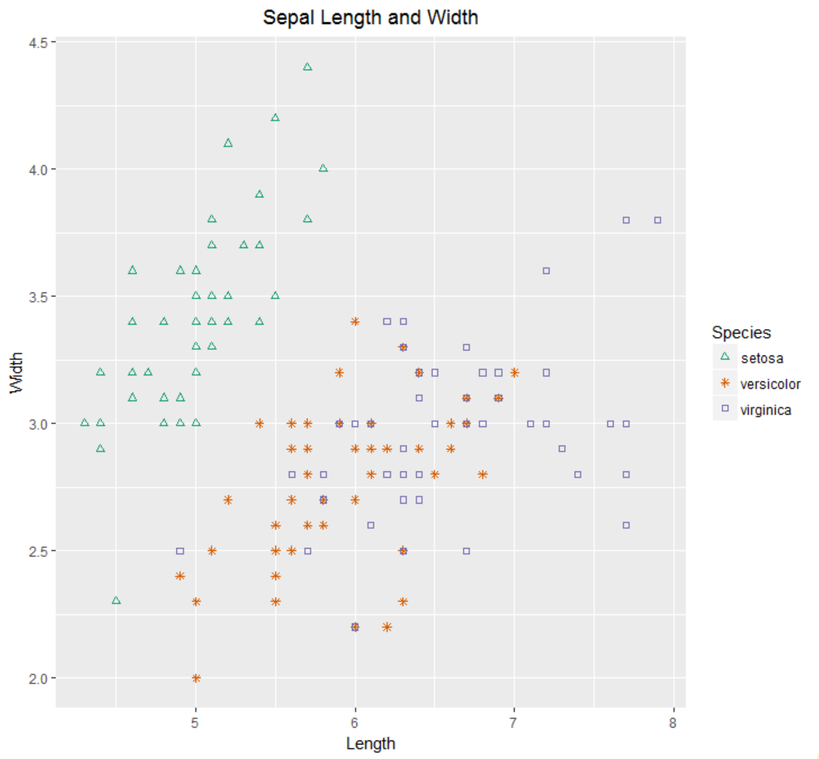


scale_colour_brewer



Scatter plot

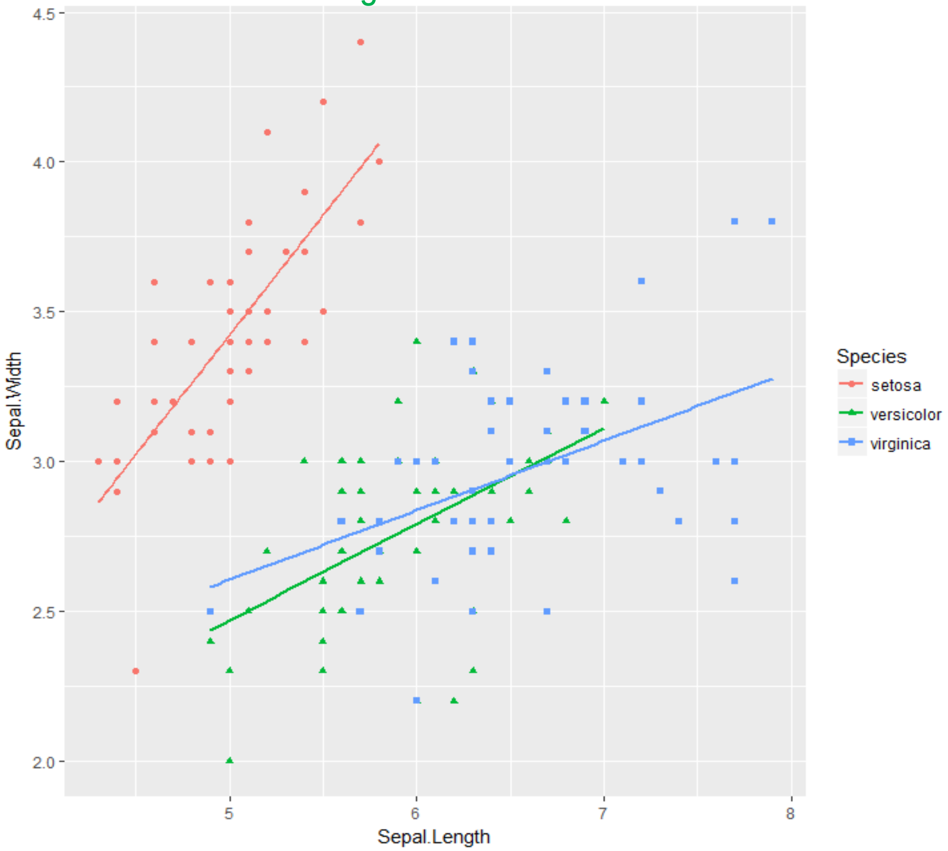
Change the points shape and colour



```
ggplot(iris, aes(x=Sepal.Length, y=Sepal.Width, shape=Species, colour=Species)) + geom_point() +
  scale_colour_brewer(palette="Dark2")+

  scale_shape_manual(values=c(2,8,0))+

  labs(x="Length",y="Width",title="Sepal Length and Width")+

  theme(plot.title = element_text(hjust = 0.5))
```

Scatter plot

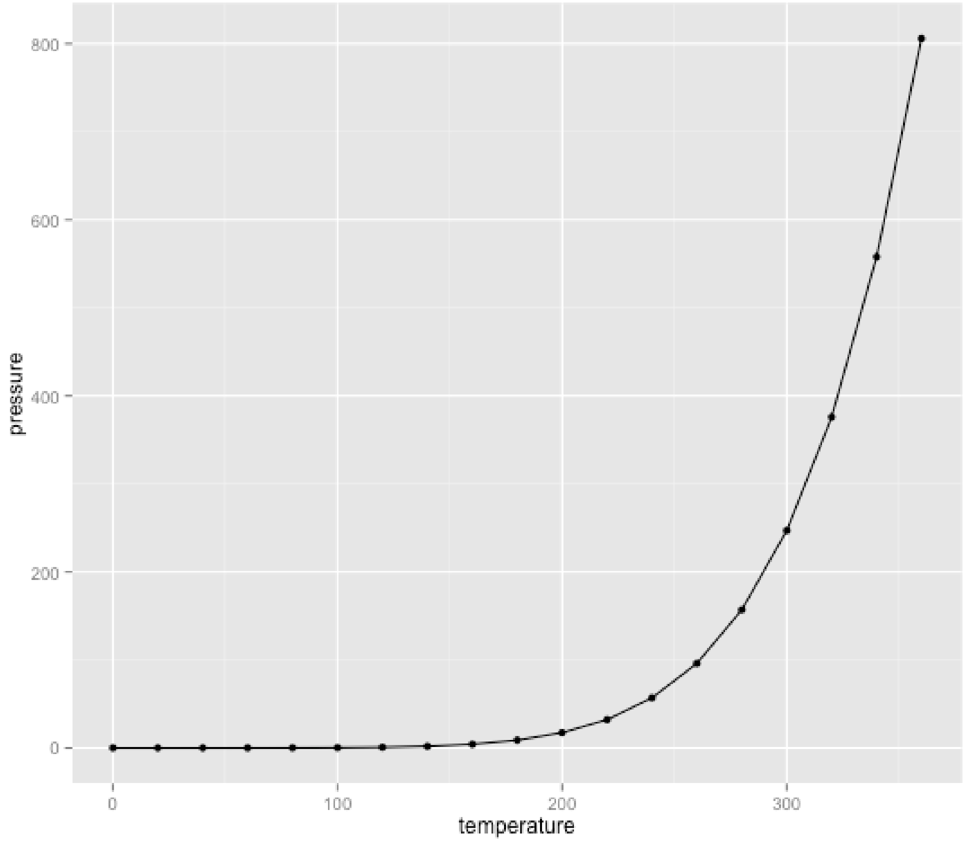
Add regression line



```
ggplot(iris, aes(x=Sepal.Length, y=Sepal.Width, shape=Species, colour=Species)) +
  geom_point() +
  geom_smooth(method=lm, se=FALSE)
```

Line Graph

the trend over time or other continuous variables

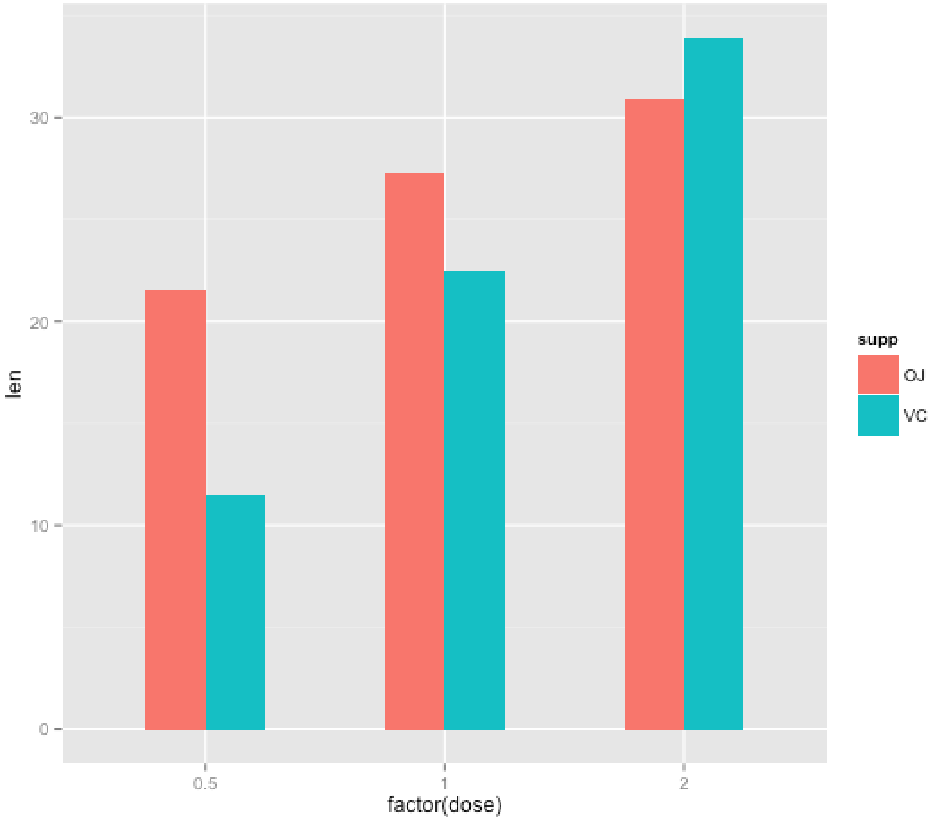


```
ggplot(pressure, aes(x=temperature, y=pressure)) + geom_line() + geom_point()
```

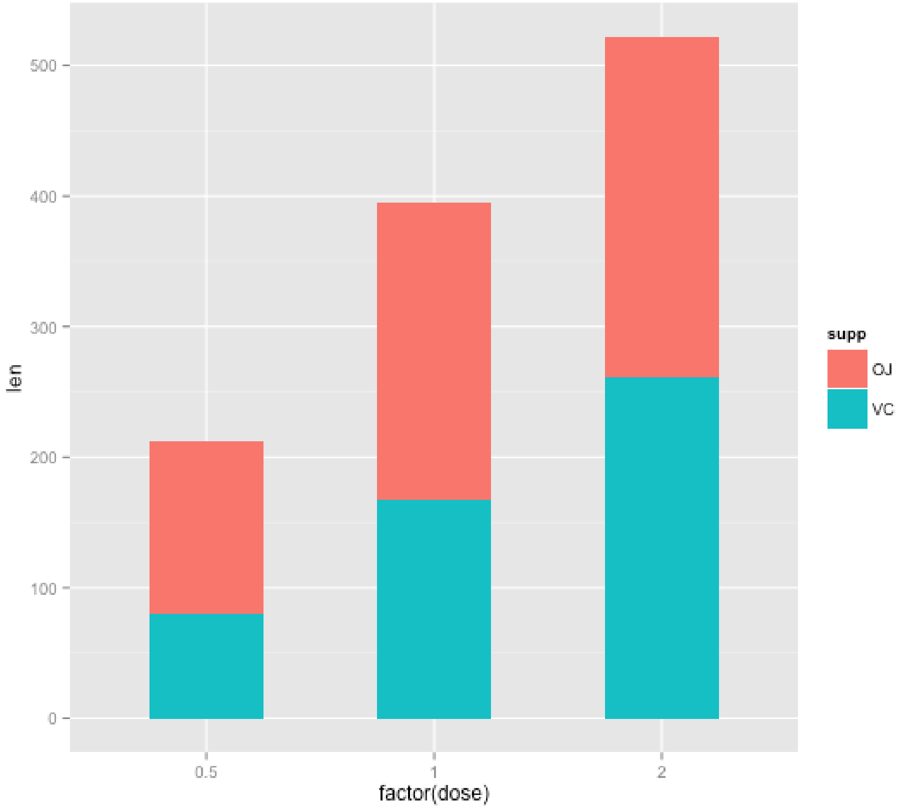
Bar Graph

display numeric values (y-axis) for different categories (x-axis)

1. Bar graph for exact value for y



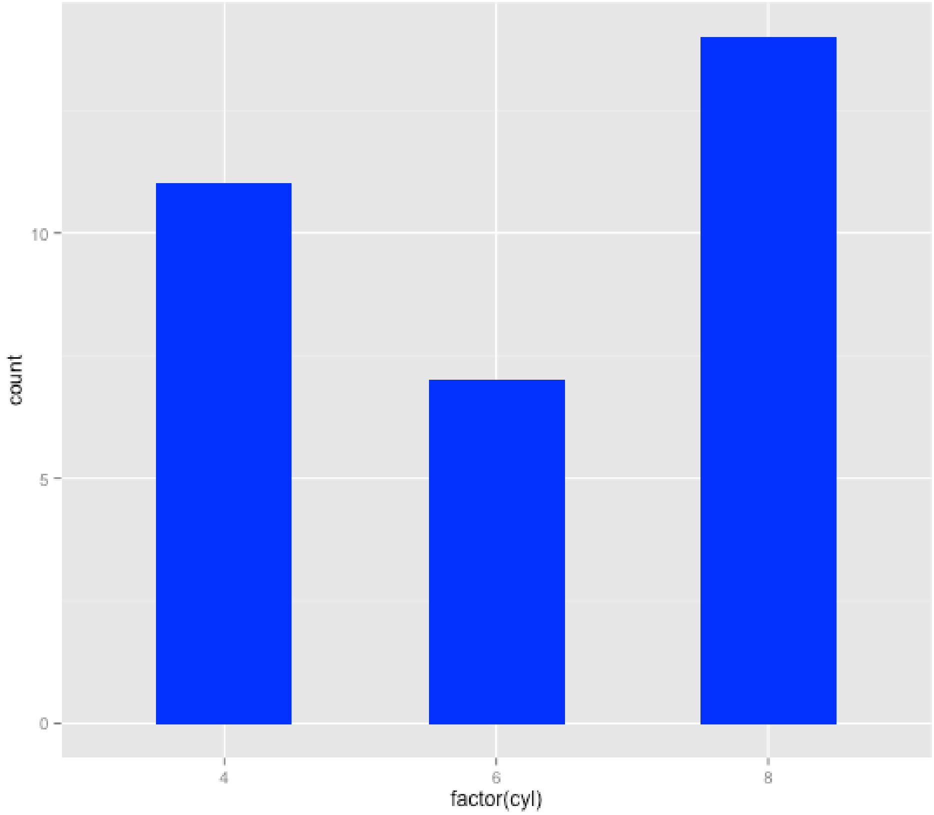
```
ggplot(ToothGrowth,aes(x=factor(dose),y=len, fill=supp))+
  geom_bar(stat="identity", position="dodge", width=0.5)
```



```
ggplot(ToothGrowth,aes(x=factor(dose),y=len, fill=supp))+
  geom_bar(stat="identity", width=0.5)
```

Bar Graph

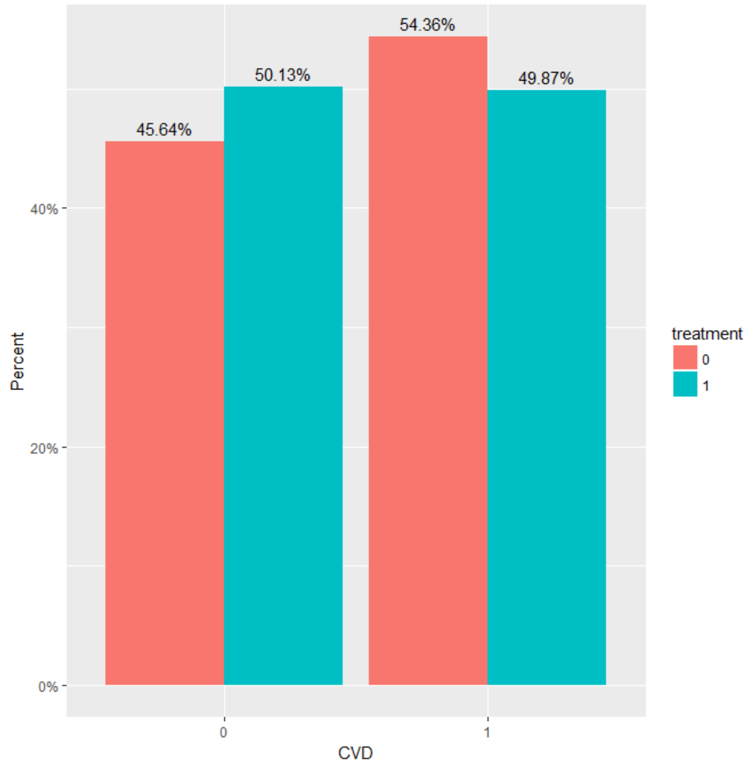
2. bar graph for counts of a categorical variable



```
ggplot(mtcars, aes(x=factor(cyl))) +
geom_bar(fill="blue",width=0.5)
```

Bar Graph

3. bar graph for percentage of a categorical variable

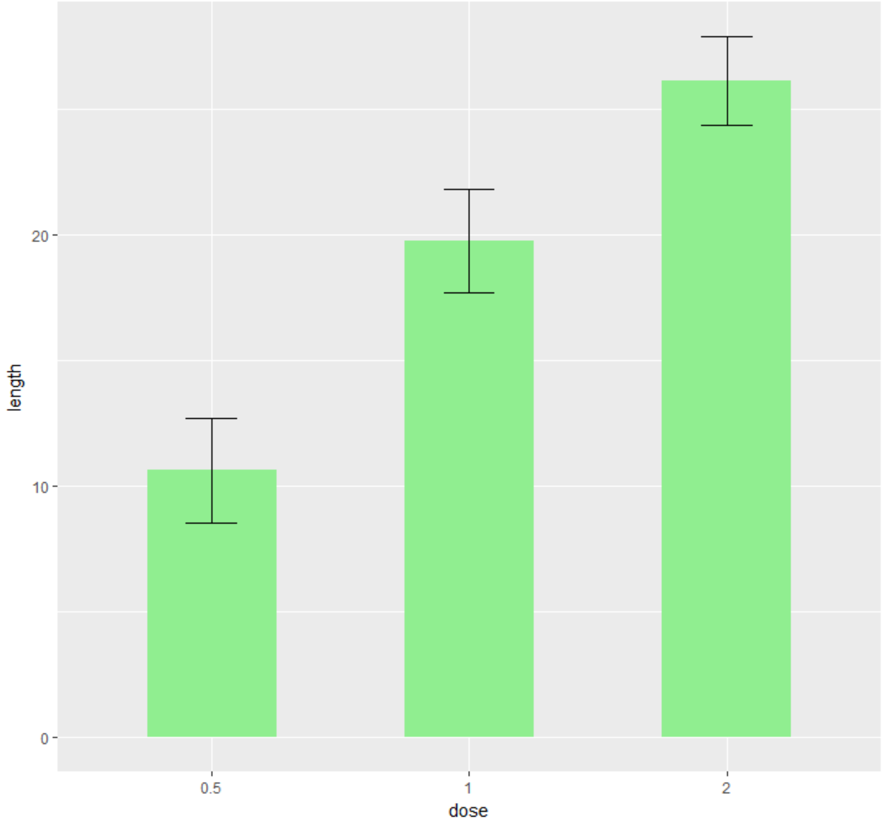


```
ggplot(dig, aes(x= factor(CVD),  group=factor(TRTMT))) +
  geom_bar(aes(y = ..prop.., fill = factor(TRTMT)), stat="count”, position="dodge")+
  geom_text(aes(label = scales::percent(..prop..), y= ..prop.. ),
           stat= "count",position=position_dodge(0.9), vjust = -.5) +
  labs(x="CVD",y = "Percent", fill="treatment") +
  scale_y_continuous(labels=scales::percent)
```

Data set description:
https://rdrr.io/cran/asympTest/man/DIGdata.html

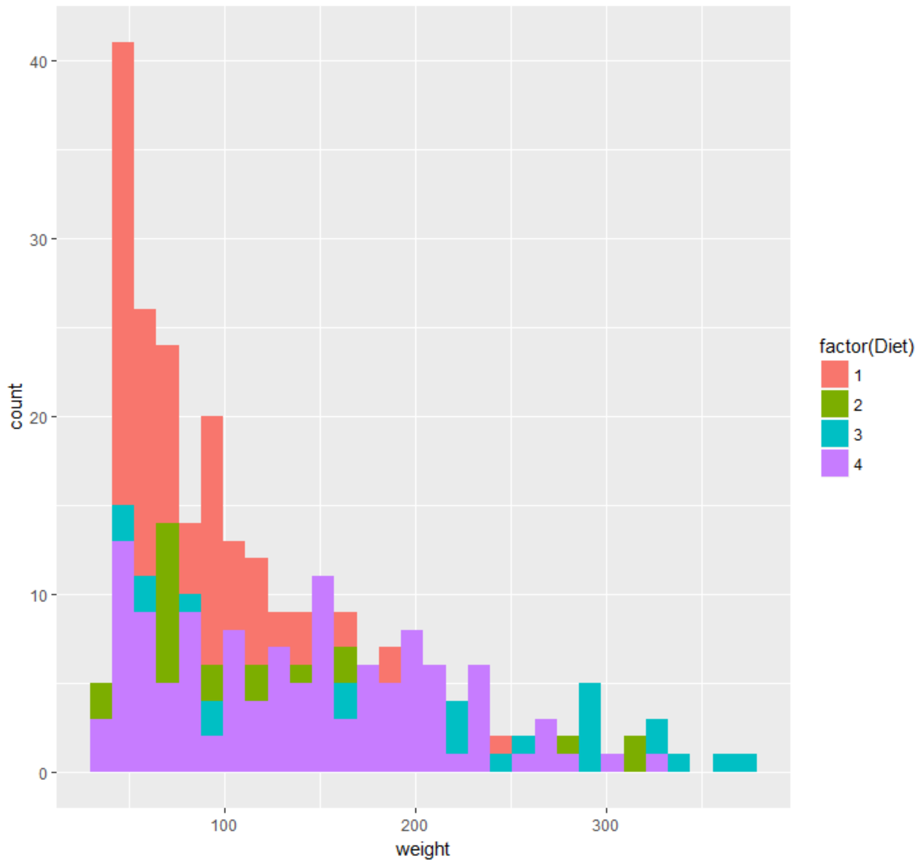
Bar Graph

4. Plot mean and error bars

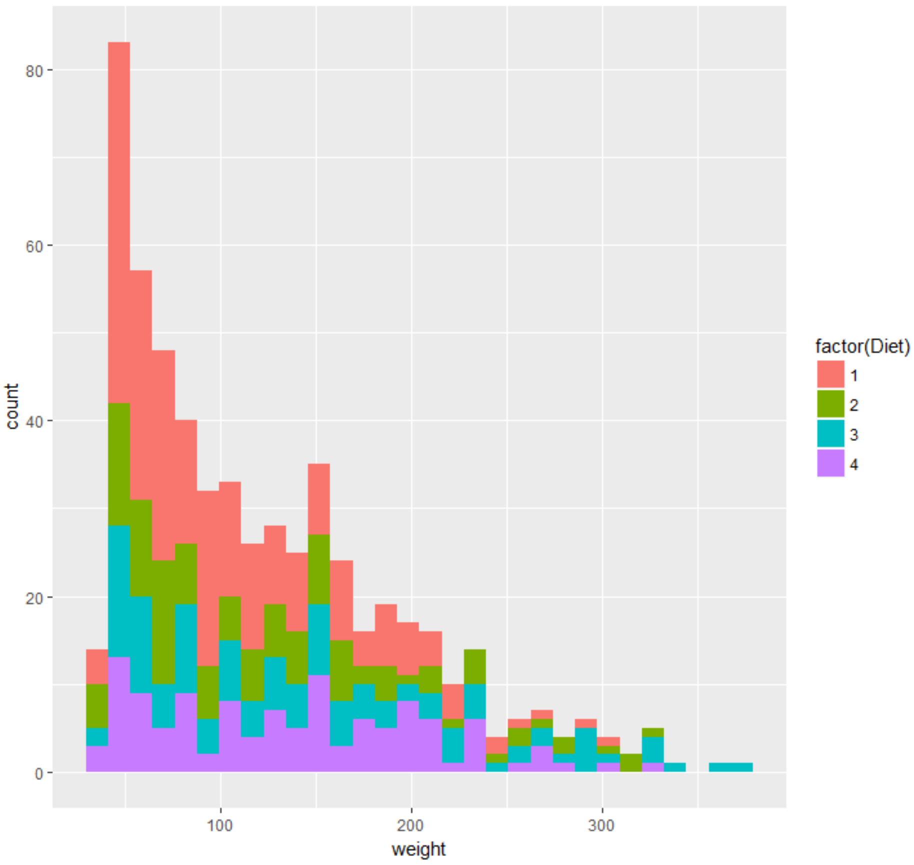


```
ggplot(ToothGrowth, aes(factor(dose), len )) +
  stat_summary(fun.y = mean, geom = "bar", width=0.5, fill="lightgreen") +
  stat_summary(fun.data = mean_cl_normal, geom = "errorbar", width=0.2)+
  labs(x="dose", y="length")
```

Visualize the distribution of data: I. histogram

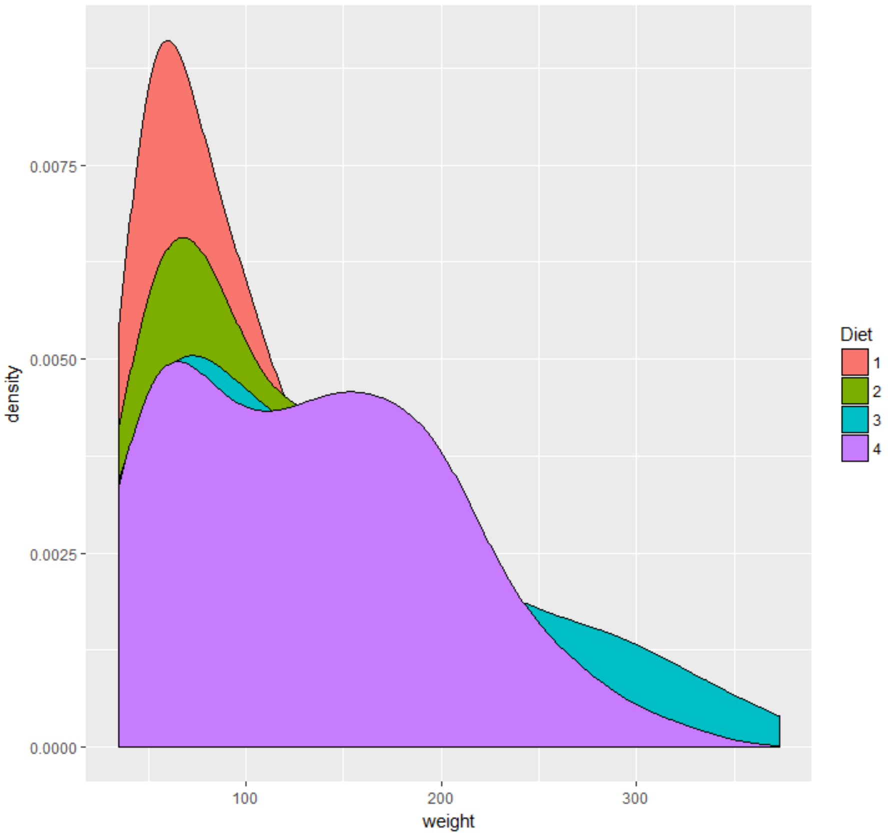


```
ggplot(ChickWeight, aes(x=weight, fill=factor(Diet))) +
  geom_histogram(position="identity")
```



```
ggplot(ChickWeight, aes(x=weight, fill=factor(Diet))) +
  geom_histogram()
```

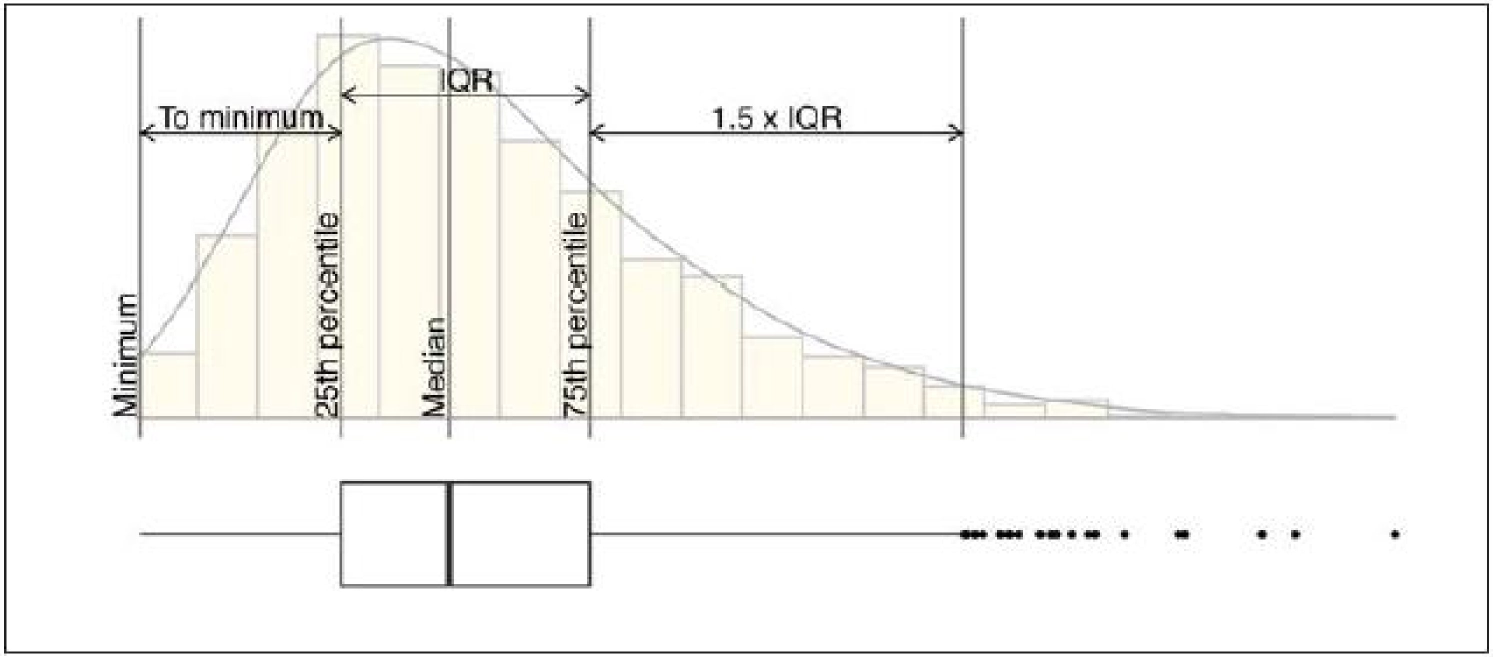
Visualize the distribution of data: II. Density Curve



```
ggplot(ChickWeight, aes(x=weight, fill=Diet)) +
geom_density()
```

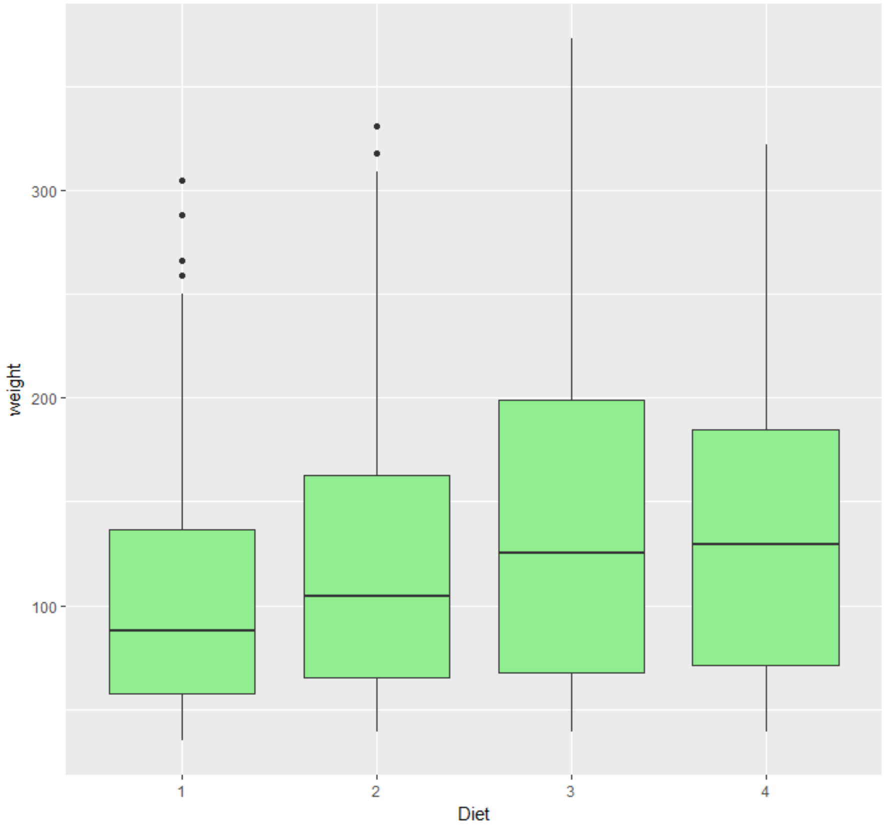
Visualize the distribution of data: III. Boxplot

Transformation among histogram, density curve and boxplot



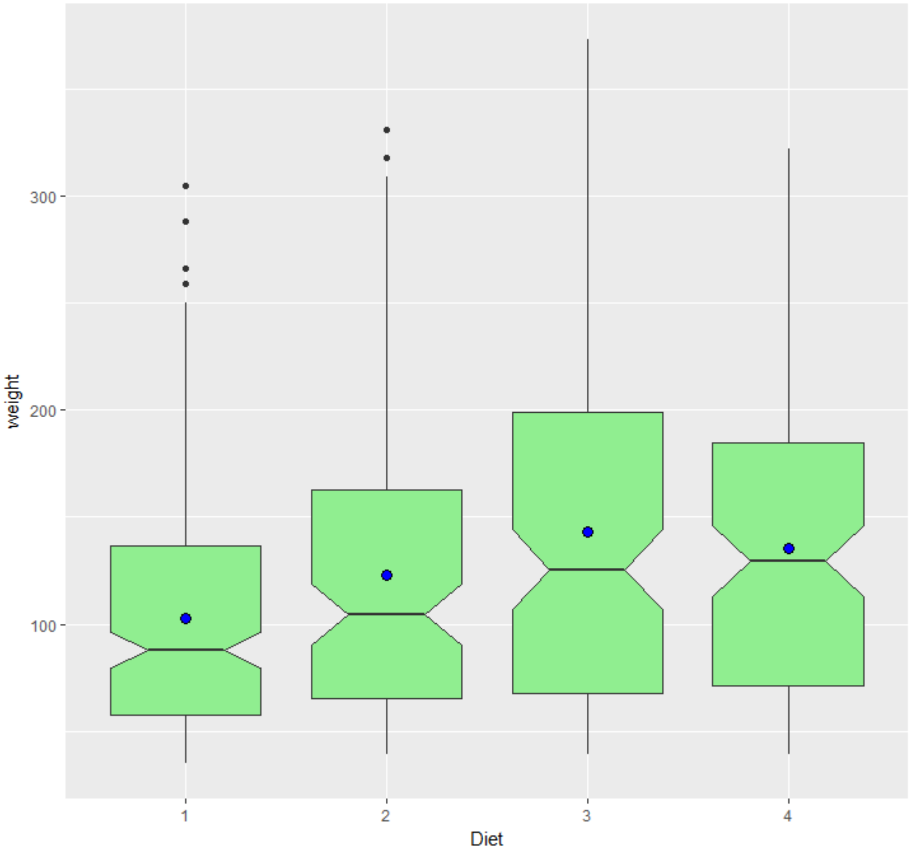
"*R Graphics Cookbook* by Winston Chang (O'Reilly). Copyright 2013 Winston Chang, 978-1-449-31695-2."

Visualize the distribution of data: III. Boxplot



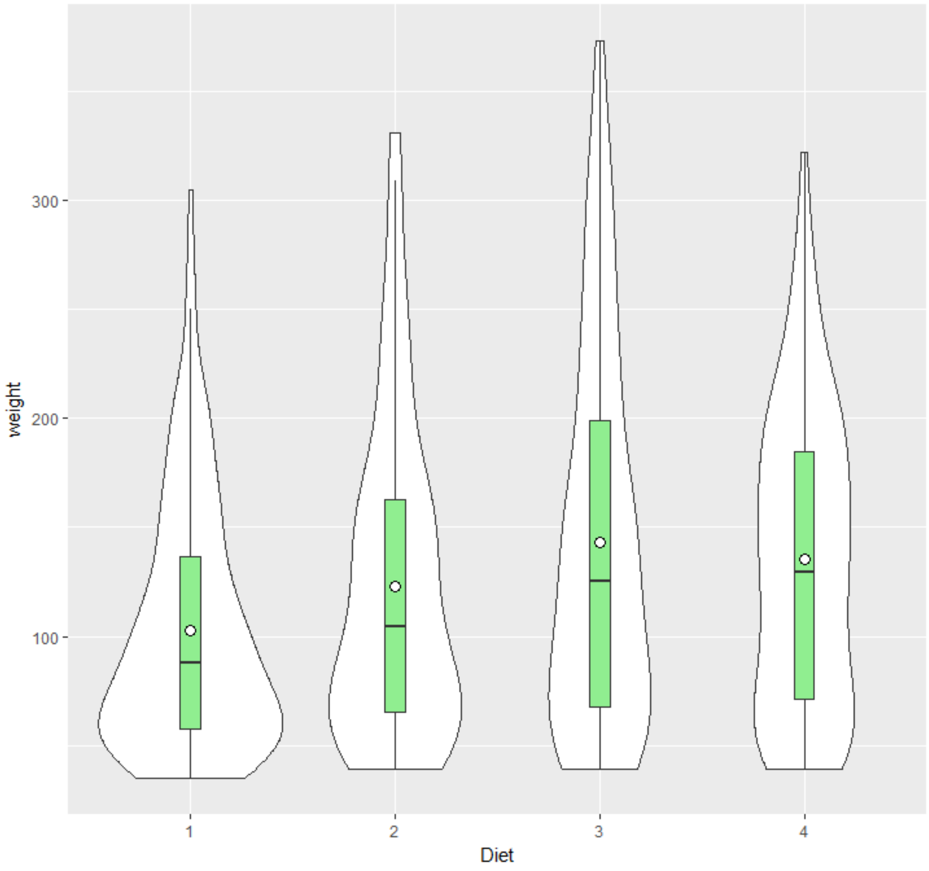
```
ggplot(ChickWeight, aes(x=Diet, y=weight)) +
geom_boxplot(fill="lightgreen")
```

Visualize the distribution of data: III. Boxplot



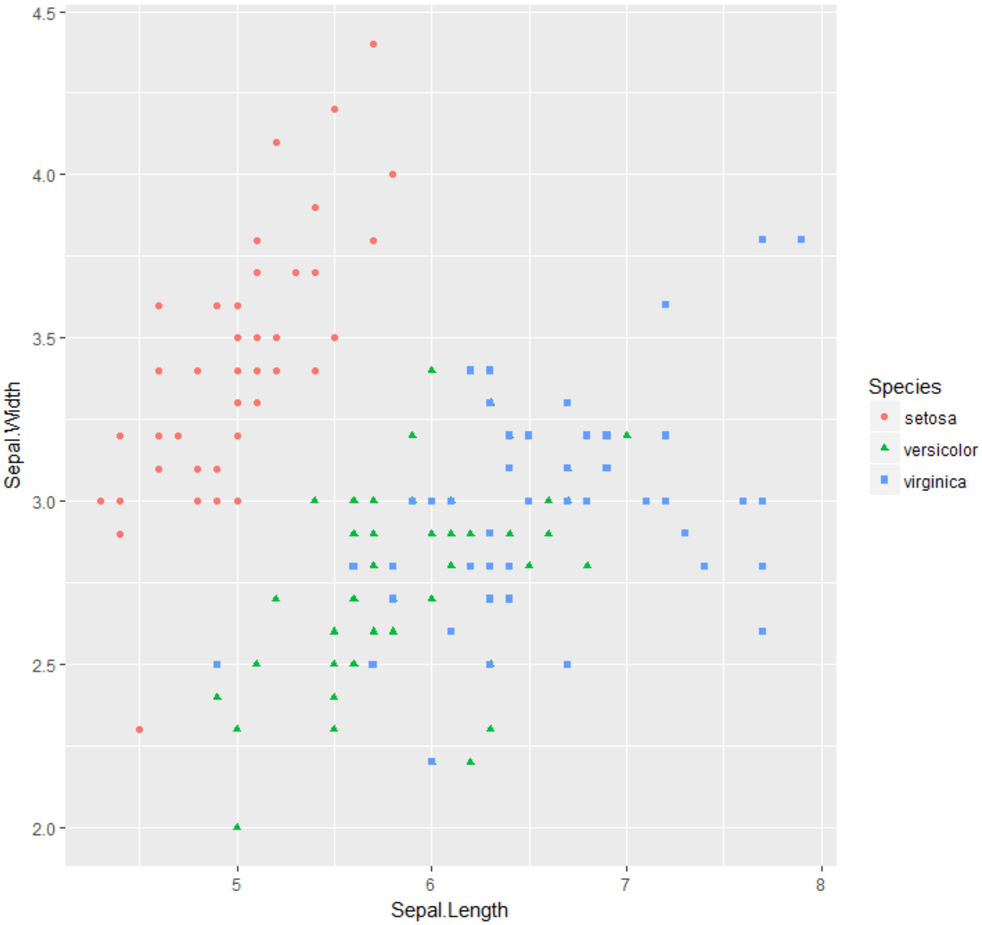
```
ggplot(ChickWeight, aes(x=Diet, y=weight)) +
geom_boxplot(fill="lightgreen",notch=TRUE)+
stat_summary(fun.y="mean",geom="point",  fill="blue", shape=21, size=3)
```

Visualize the distribution of data: III. Boxplot



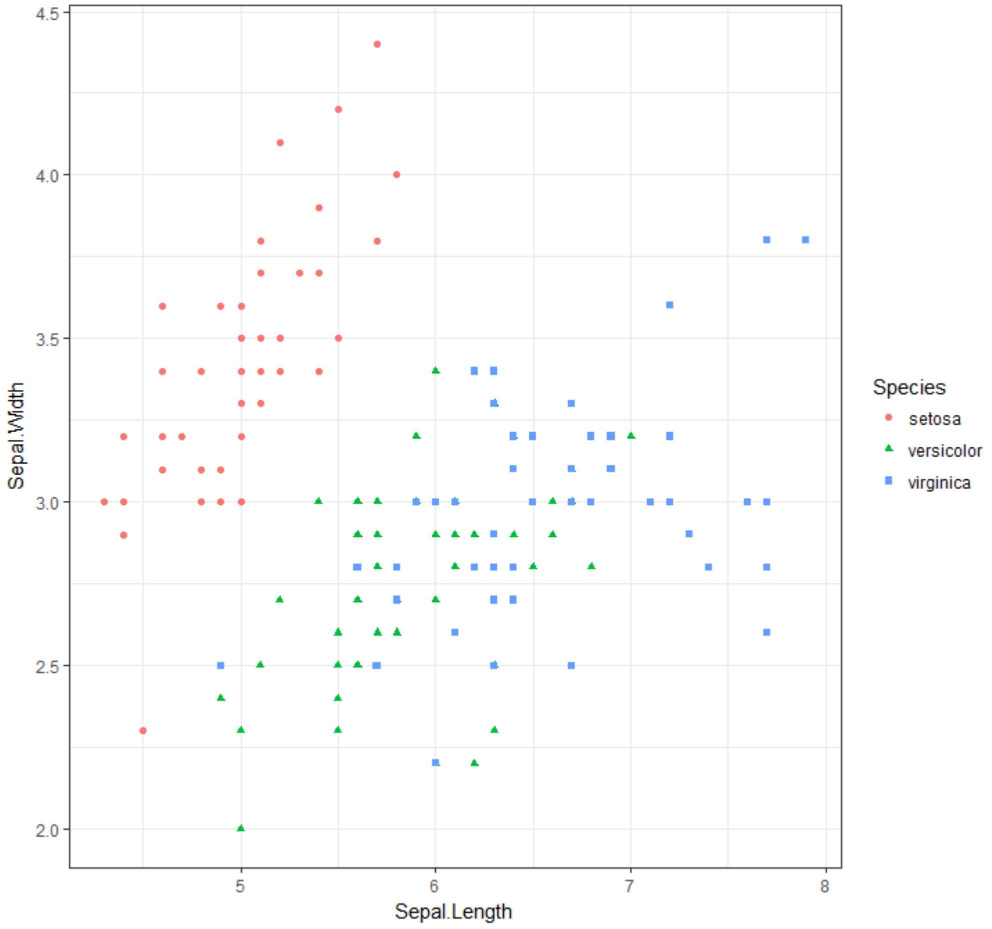
```
ggplot(ChickWeight, aes(x=Diet, y=weight))  + geom_violin() +
  geom_boxplot(width=.1, fill="lightgreen", outlier.colour=NA) +
  stat_summary(fun.y=mean, geom="point", fill="white", shape=21, size=3)
```

Change the overall appearance of the graph



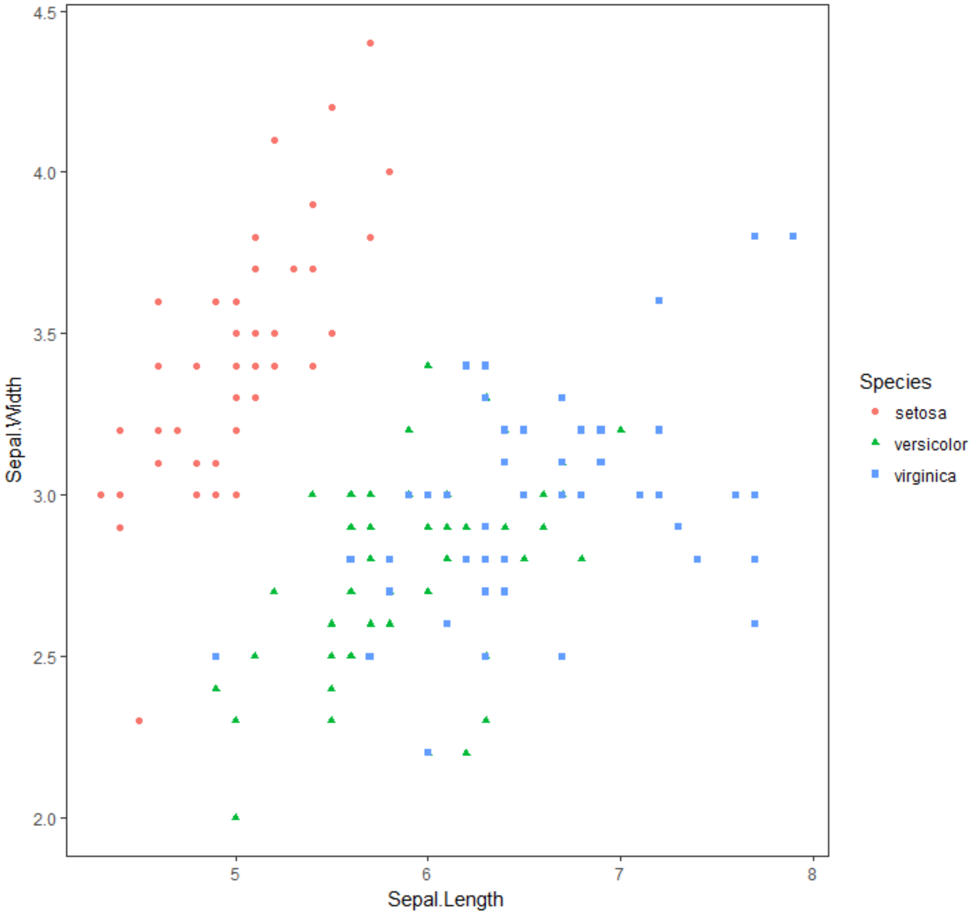
```
bp <- ggplot(iris, aes(x=Sepal.Length, y=Sepal.Width, color=Species, shape=Species))+
        geom_point()
bp
```

Set background to be black and white



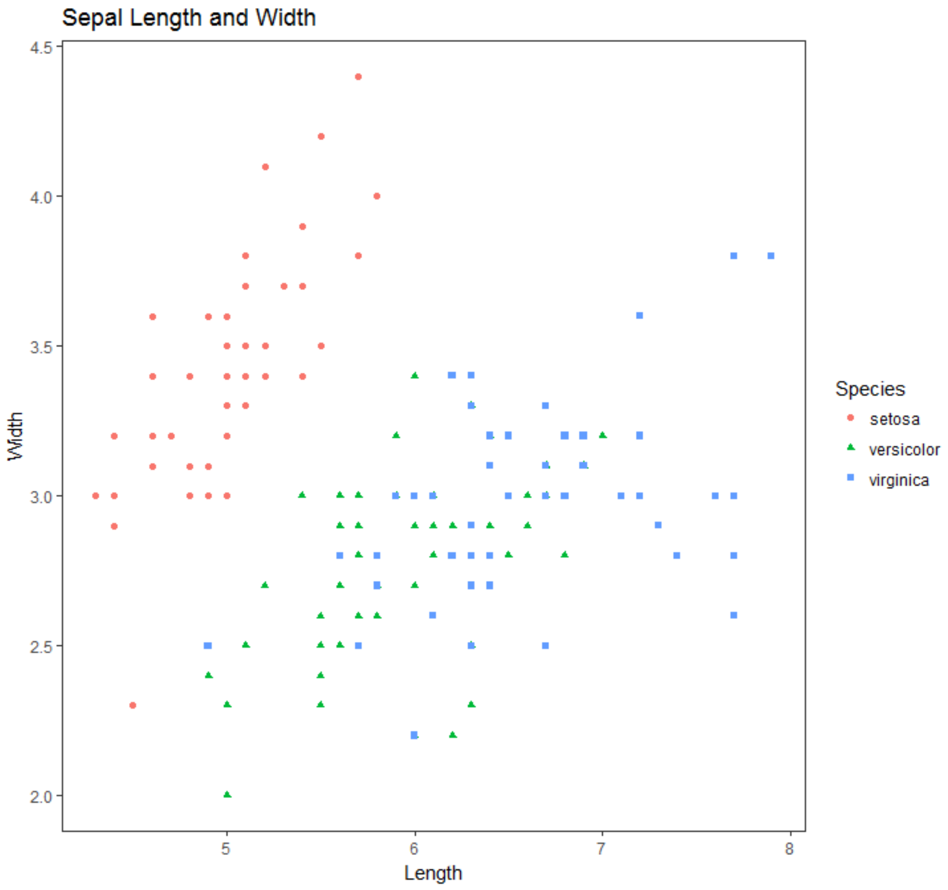
```
bp1<-bp + theme_bw()
bp1
```

Remove grid lines



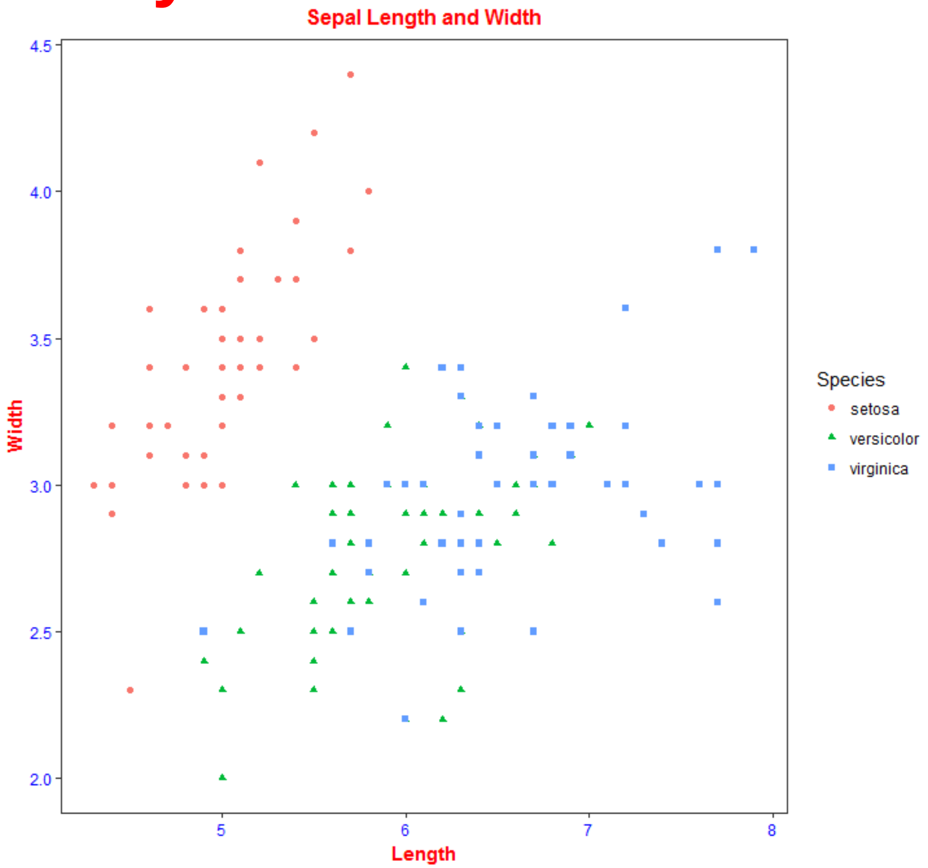
```
bp2<-bp1 + theme(panel.grid.major = element_blank(),panel.grid.minor = element_blank())
bp2
```

Add Labels



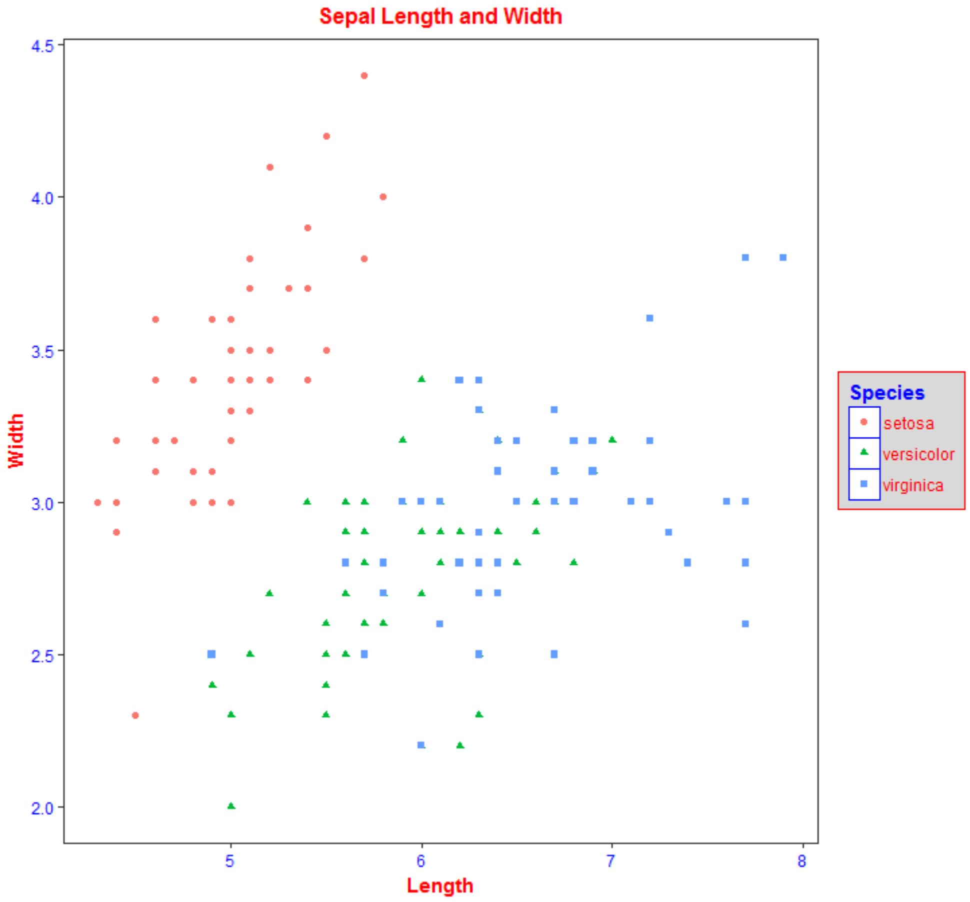
```
bp3<- bp2+labs(x="Length",y="Width",title="Sepal Length and Width")
bp3
```

Modify title and axis labels



```
bp4<- bp3 +
  theme(axis.title.x = element_text(colour="red", size=11,face="bold"),
        axis.text.x  = element_text(colour="blue"),
        axis.title.y  = element_text(colour="red", size=11,face="bold", angle = 90),
        axis.text.y  = element_text(colour="blue"),
        plot.title = element_text(colour="red", size=12, face="bold", hjust=0.5))
bp4
```

Modify legend



```
bp4 +
 theme(legend.background = element_rect(fill="grey85", colour="red", size=0.2),
       legend.title = element_text(colour="blue", face="bold", size=11),
       legend.text = element_text(colour="red"),
       legend.key = element_rect(colour="blue", size=0.2))
```

Simplified code (plotting for single group)

Default tools in R		ggplot
plot (x, y, type=“p”)	Scatter plot	ggplot(data, aes(x= , y=,)) + geom_point()
plot (x, y, type=“l”)	Line graph	ggplot(data, aes(x= , y=,)) + geom_line()
barplot(x=table(x))	Bar graph	ggplot(data, aes(x=factor())) + geom_bar()
hist(x)	Histogram	ggplot(data, aes(x=)) + geom_histogram()
boxplot(data=data, y~x)	Boxplot	ggplot(data, aes(x=factor(),y=)) + geom_boxplot()

Simplified code (plotting for multiple groups)

Scatter plot

```
ggplot(data, aes(x= , y=, shape=factor(group), colour=factor(group))) +
  geom_point()
```

Bar graph

```
ggplot(data, aes(x=factor(), fill=factor(group))) +
  geom_bar(postion=“dodge”)
```

Histogram

```
ggplot(data, aes(x=,fill=factor(group))) +
  geom_histogram(position =“identity”)
```

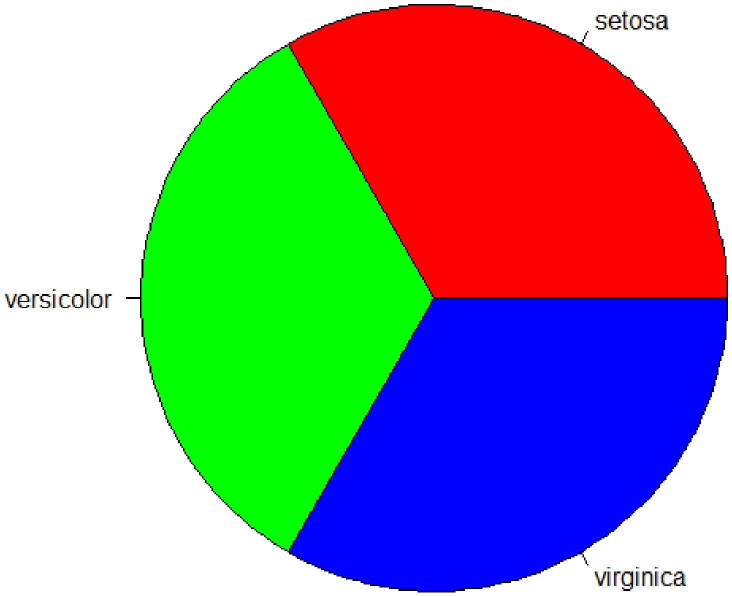
Density Curve

```
ggplot(data, aes(x=,fill=factor(group))) +
  geom_density()
```

Boxplot

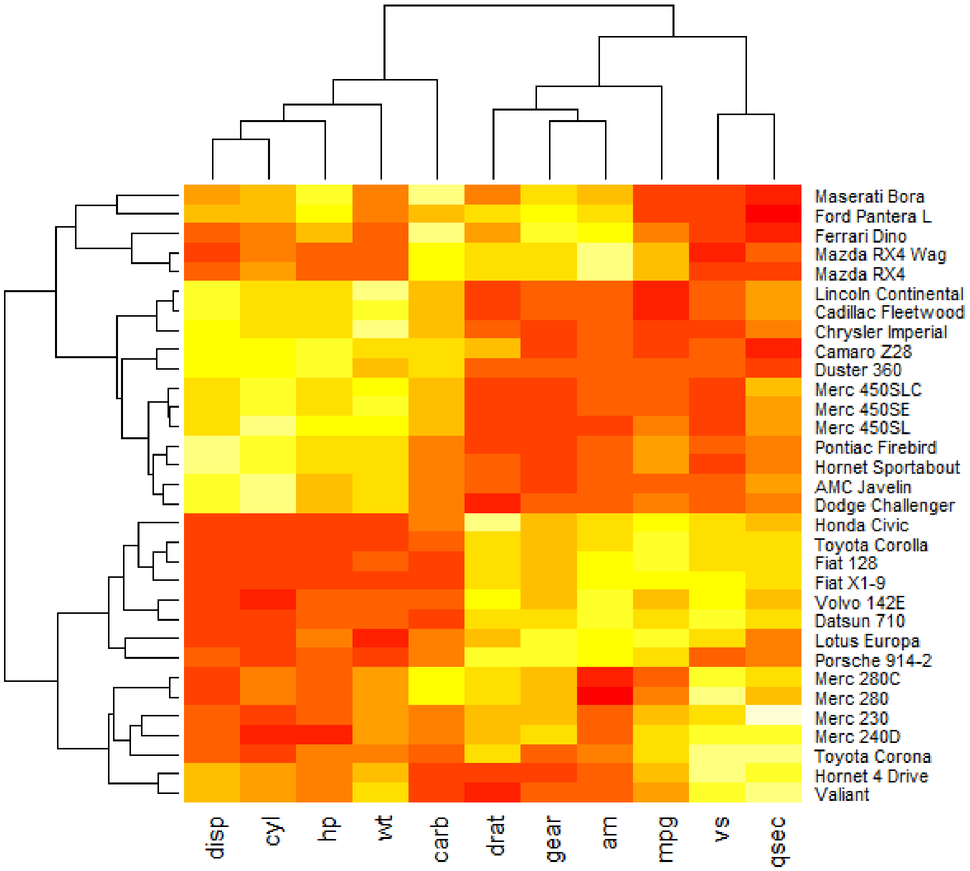
```
ggplot(data, aes(x=factor(),y=, fill=factor(group))) +
  geom_boxplot()
```

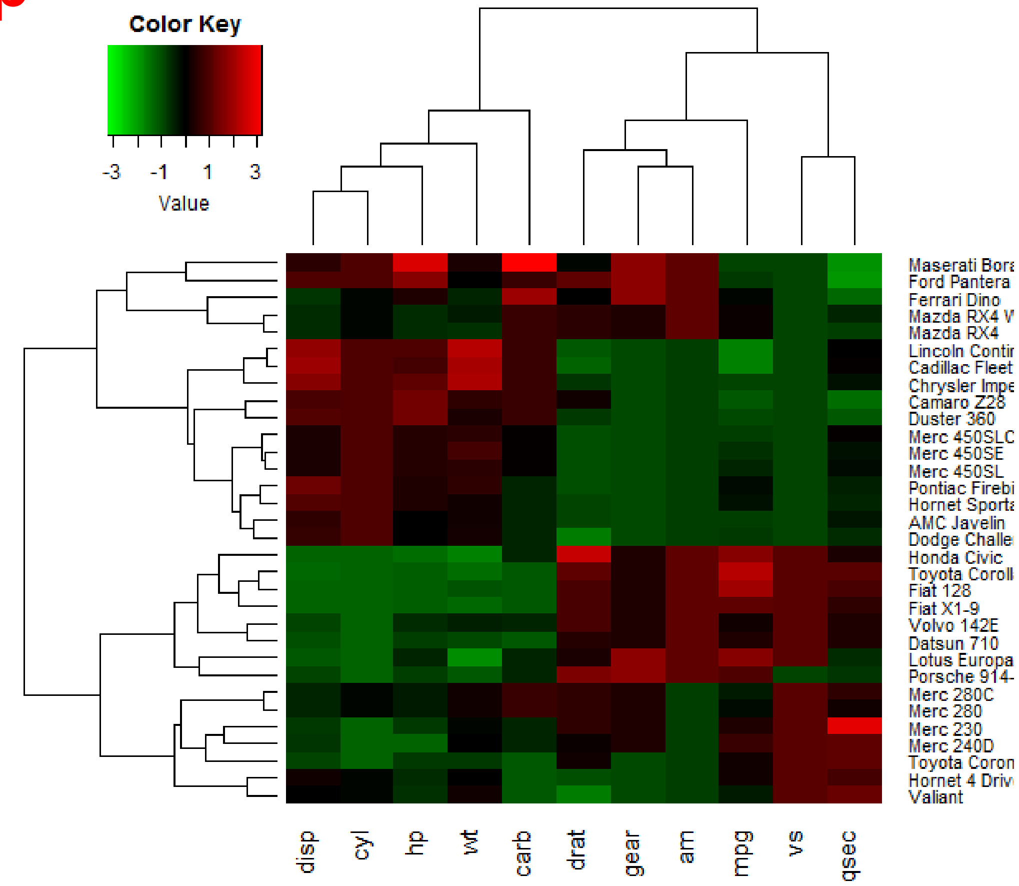
Pie chart



```
pie(table(iris$Species), col=rainbow(3))
```

Heatmap



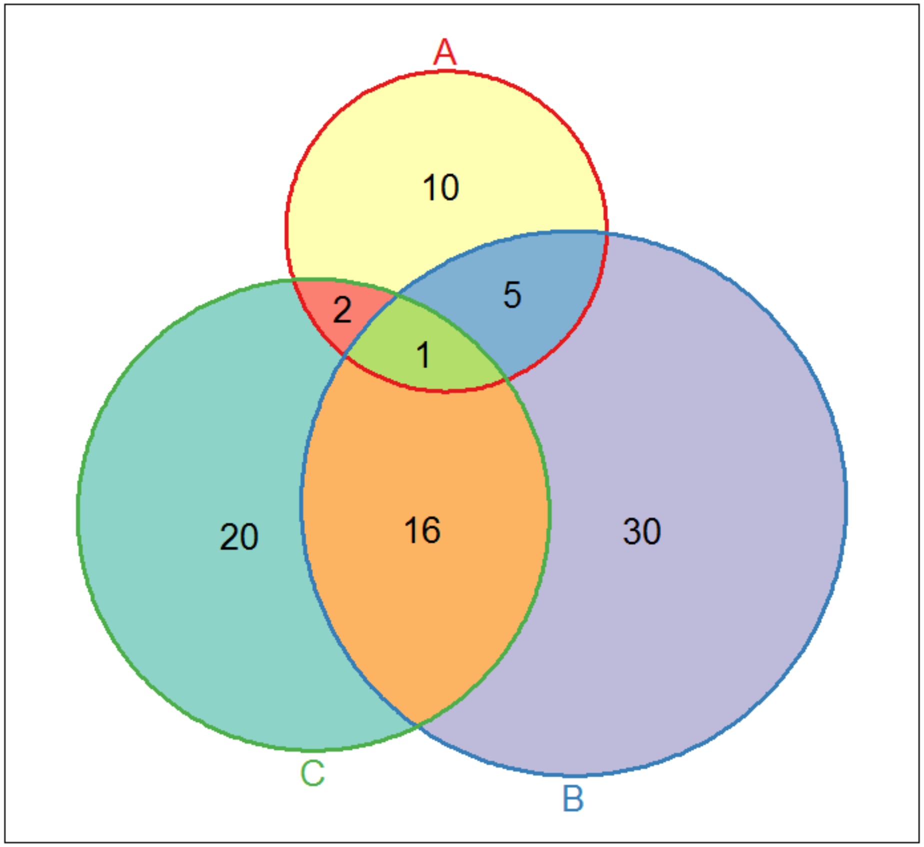


```
data <- as.matrix(scale(mtcars))
Heatmap(data)
```

```
data <- as.matrix(scale(mtcars))
library(gplots)
heatmap.2(data,col=greenred(100),trace="none",density.info="none")
```

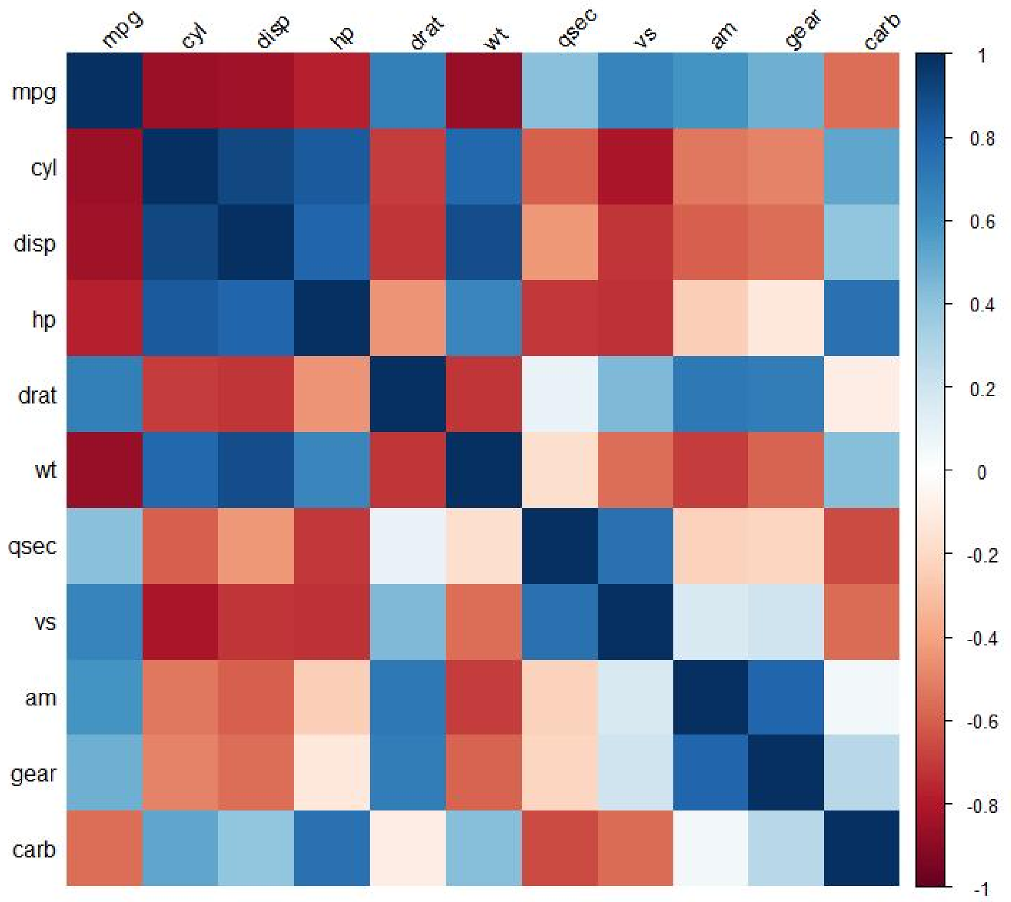
Powerful online tools for heatmaps: https://software.broadinstitute.org/morpheus/

Venn Diagram



```
library(Vennerable)
V <- Venn(SetNames=c('A','B','C'),Weight=c(0,10,30,5,20,2,16,1))
plot(V, doWeights=TRUE,type='circles')
```

Correlation Plot



```
mcor<-cor(mtcars)
library(corrplot)
corrplot(mcor, method="shade", shade.col=NA, tl.col="black", tl.srt=45)
```

Output for publication or presentation

Output to PNG:

```
png ( filename = " Scatterplot .png", width = 480 ,height = 480)
plot ( pressure )
dev .off ()
```

Output to PDF:

```
pdf ( filename = " Scatterplot .pdf", width = 4, height= 4)
plot ( pressure )
dev .off ()
```

Note: dev .off () is to let R know you’re finished with plotting commands and it can output the file.

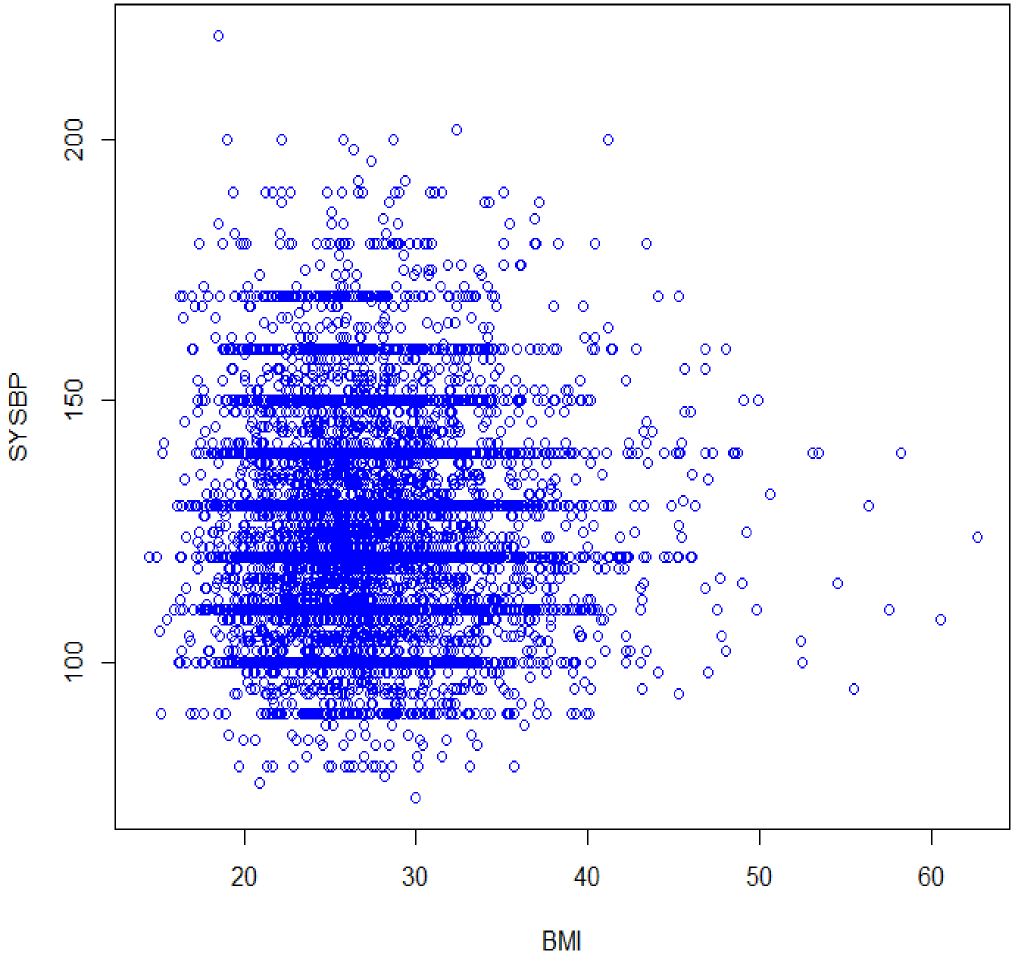
Practice Session

1. An R script that includes all the code covered in this presentation is provided with detailed step-wise annotation and you will go over all the plots during the first half of the practice session.

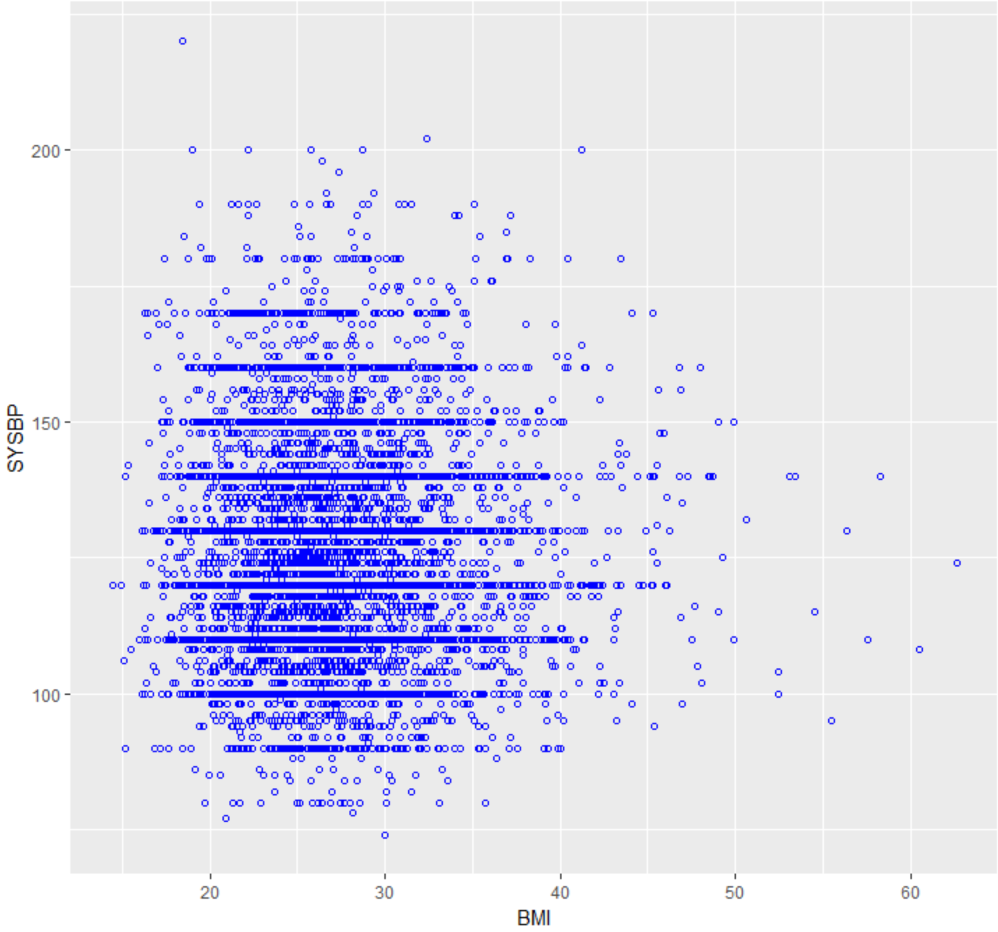
2. You will try to answer a list of five questions and draw plots to visualize the practicing clinical trial data set

Data set description: https://rdrr.io/cran/asympTest/man/DIGdata.html

Q1: Check the relationship between BMI and Systolic BP

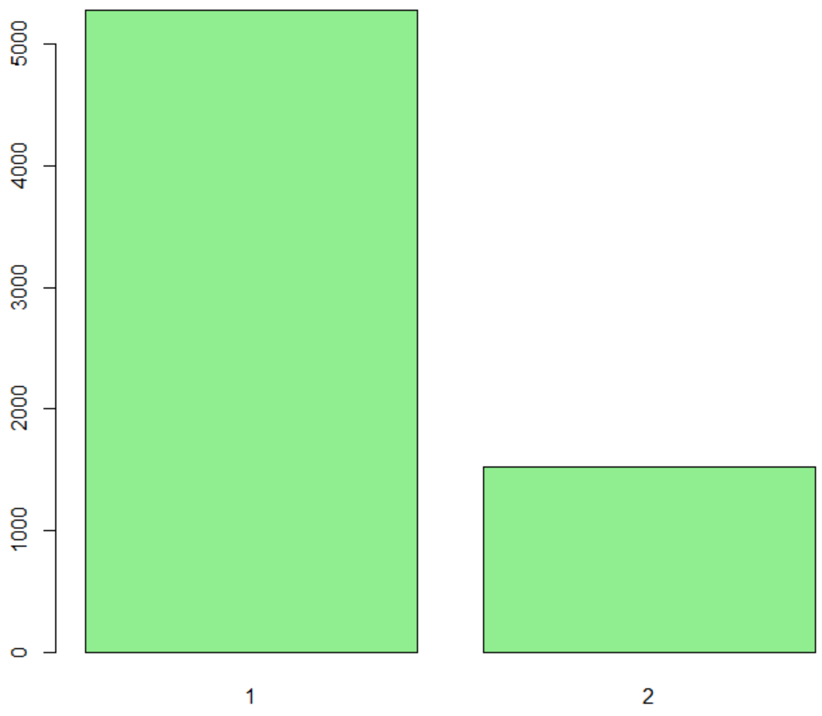


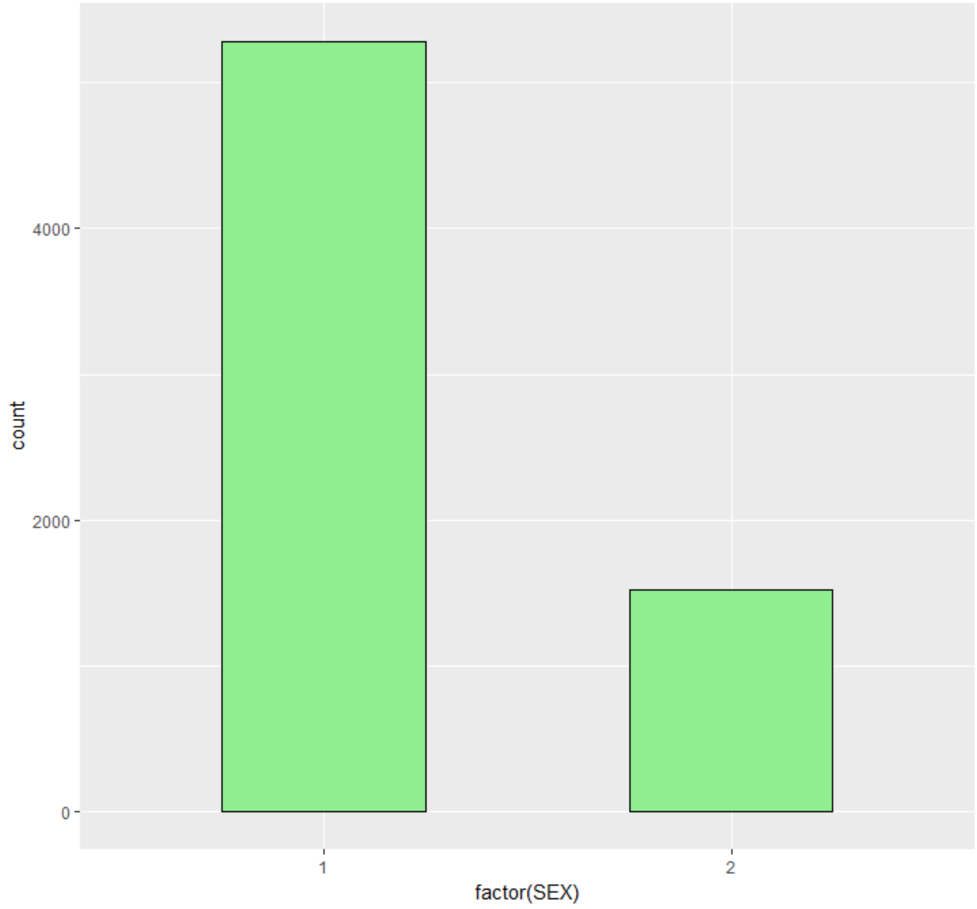
```
plot(dig$BMI,dig$SYSBP, xlab="BMI", ylab="SYSBP", col="blue")
```



```
ggplot(dig, aes(x=BMI, y=SYSBP)) +
 geom_point(colour="blue", shape=21)
```

Q2: Plot the number of patients for different SEX group

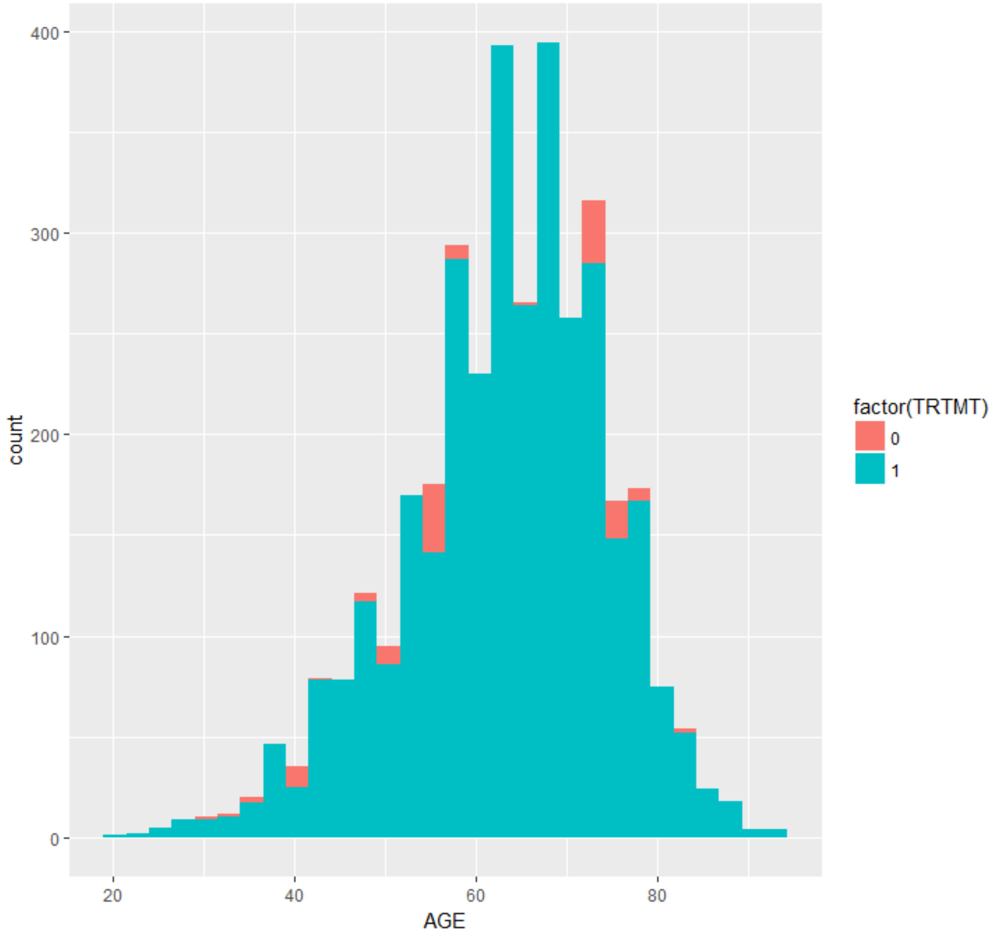




barplot(table(dig$SEX), col="lightgreen")

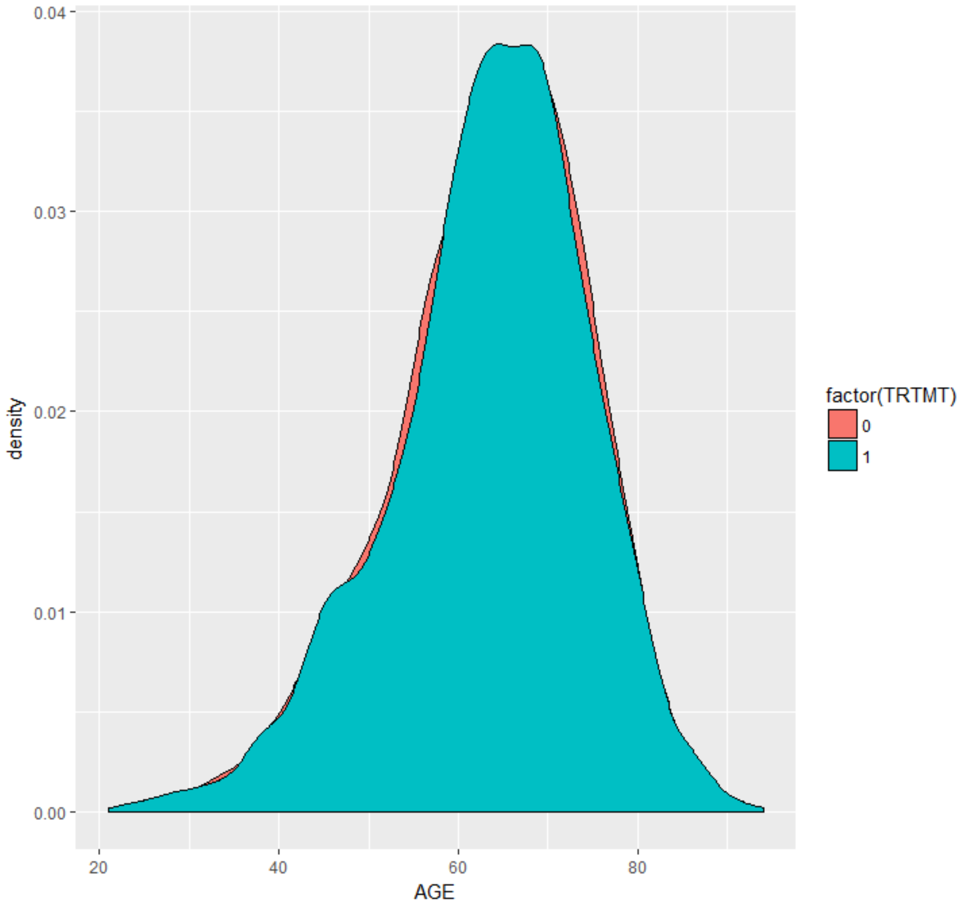
ggplot(dig, aes(x=factor(SEX))) +
geom_bar(colour="black", fill="lightgreen", width=0.5)

Q3: Use ggplot to check the distribution of age in different treatment group using three different types of plots - histogram



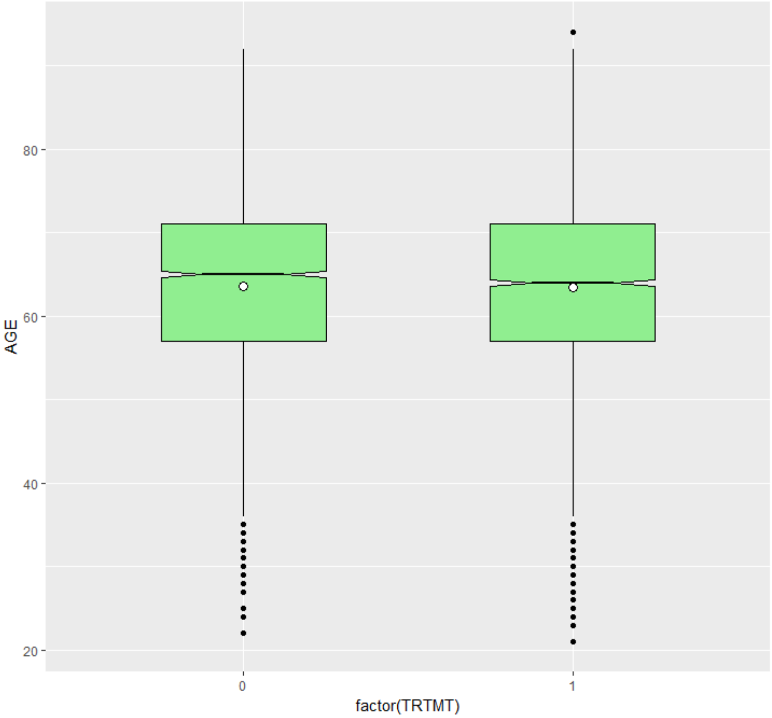
```
ggplot(dig, aes(x=AGE, fill=factor(TRTMT))) +
geom_histogram(position="identity")
```

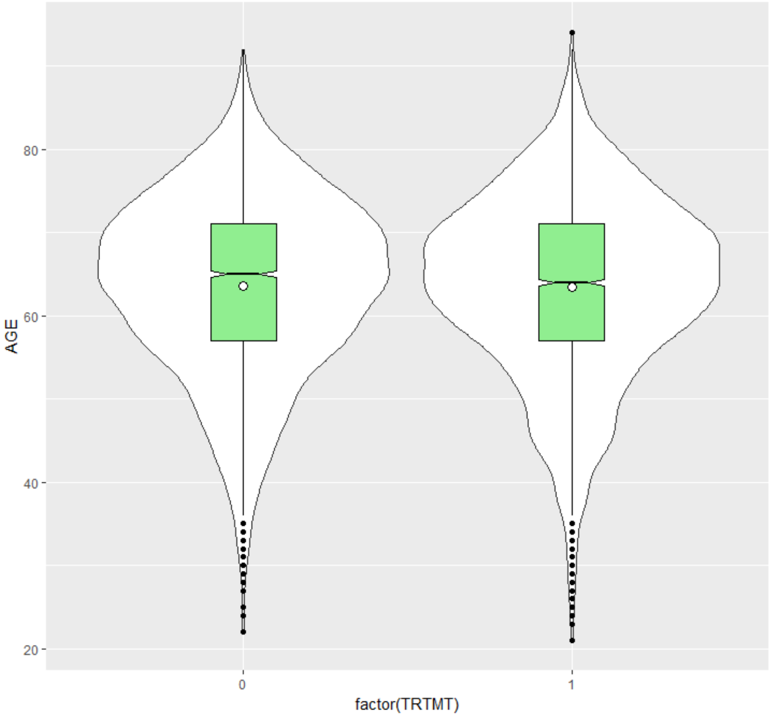
Q3: use ggplot to check the distribution of age in different treatment group using three different types of plots – density curve



```
ggplot(dig, aes(x=AGE, fill=factor(TRTMT))) +
geom_density()
```

Q3: use ggplot to check the distribution of age in different treatment group using three different types of plots – boxplot

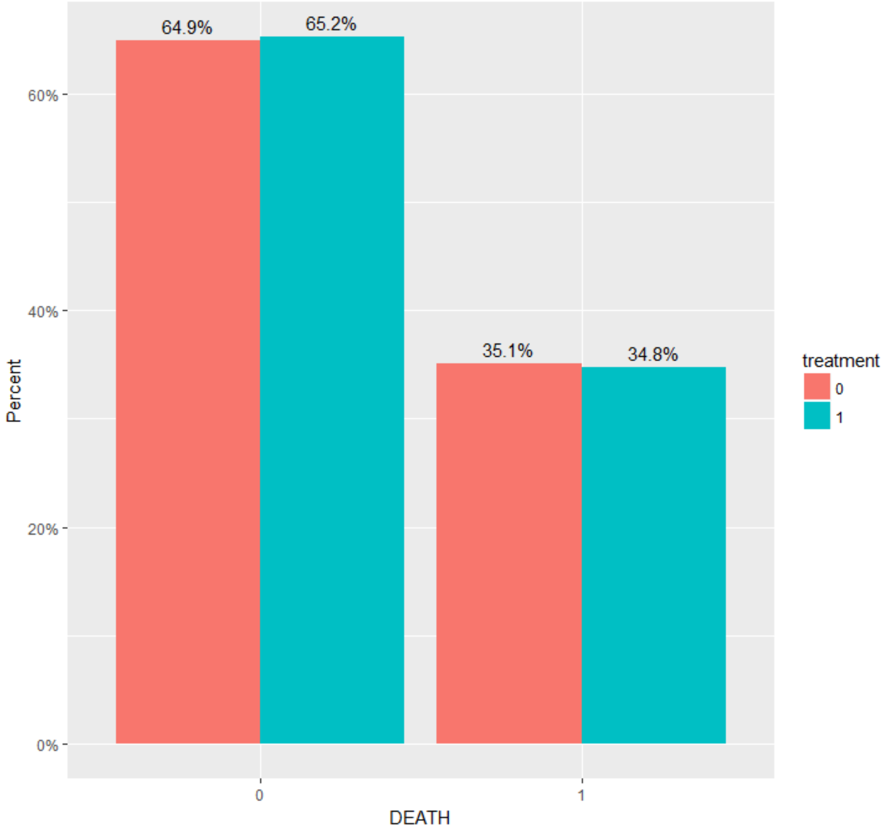




```
ggplot(dig, aes(x=factor(TRTMT), y=AGE)) +
geom_boxplot(notch=TRUE, width=0.5, colour="black", fill="lightgreen") +
  stat_summary(fun.y="mean",geom="point",  fill="white", shape=21, size=3)
```

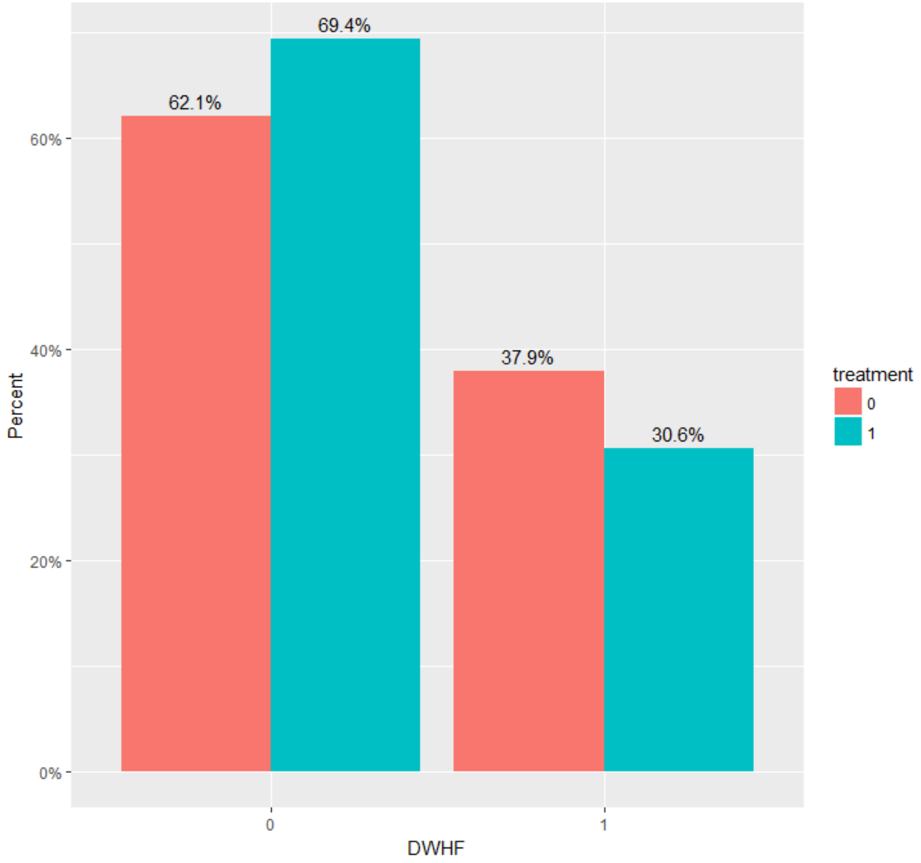
```
ggplot(dig, aes(x=factor(TRTMT), y=AGE))  + geom_violin() +
geom_boxplot(notch=TRUE, width=0.2, colour="black", fill="lightgreen")+
  stat_summary(fun.y="mean",geom="point",  fill="white", shape=21, size=3)
```

Q4. A. plot the percentage of death in different treatment group



```
ggplot(dig, aes(x= factor(DEATH),  group=factor(TRTMT))) +
  geom_bar(aes(y = ..prop.., fill = factor(TRTMT)), stat="count", position="dodge")+
  geom_text(aes( label = scales::percent(..prop..), y= ..prop.. ), stat= "count",position=position_dodge(0.9),
vjust = -.5) +
  labs(x="DEATH",y = "Percent", fill="treatment") +
  scale_y_continuous(labels=scales::percent)
```

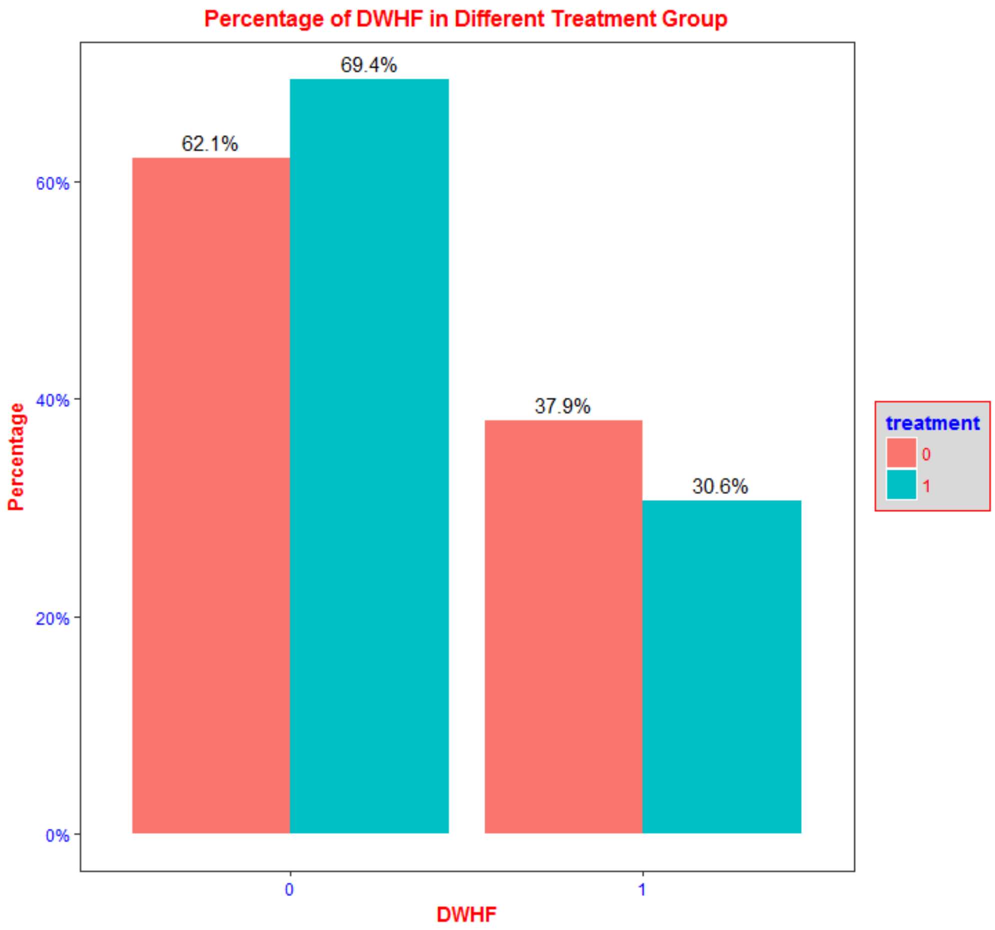
Q4. B. plot the percentage of death attributed to worsening heart failure



```
ggplot(dig, aes(x= factor(DWHF),  group=factor(TRTMT))) +
  geom_bar(aes(y = ..prop.., fill = factor(TRTMT)), stat="count", position="dodge")+
  geom_text(aes( label = scales::percent(..prop..), y= ..prop.. ), stat= "count",position=position_dodge(0.9),
vjust = -.5) +
  labs(x="DWHF",y = "Percent", fill="treatment") +
  scale_y_continuous(labels=scales::percent)
```

Q5 take plot from Q4b, try to polish the graph

```
Q4B<-ggplot(dig, aes(x= factor(DWHF),  group=factor(TRTMT))) +
  geom_bar(aes(y = ..prop.., fill = factor(TRTMT)), stat="count", position="dodge")+
  geom_text(aes( label = scales::percent(..prop..), y= ..prop.. ), stat=
"count",position=position_dodge(0.9),  vjust = -.5) +
  labs(x="DWHF",y = "Percent", fill="treatment") +
  scale_y_continuous(labels=scales::percent)
Q4B+
  theme_bw()+
  theme(panel.grid.major = element_blank(),panel.grid.minor = element_blank())+
  labs(x="DWHF", y="Percentage", title="Percentage of DWHF in Different Treatment
Group") +
  theme(axis.title.x = element_text(colour="red", size=11,face="bold"),
        axis.text.x = element_text(colour="blue"),
        axis.title.y = element_text(colour="red", size=11,face="bold", angle = 90),
        axis.text.y = element_text(colour="blue"),
        plot.title = element_text(colour="red", size=12, face="bold", hjust=0.5)) +
  theme(legend.background = element_rect(fill="grey85", colour="red", size=0.2),
        legend.title = element_text(colour="blue", face="bold", size=11),
        legend.text = element_text(colour="red"))
```


Percentage of DWHF in Different Treatment Group
Percentage
DWHF
treatment