### HW5\_IS457\_8

# Part I: Plot arguments (4 pts)

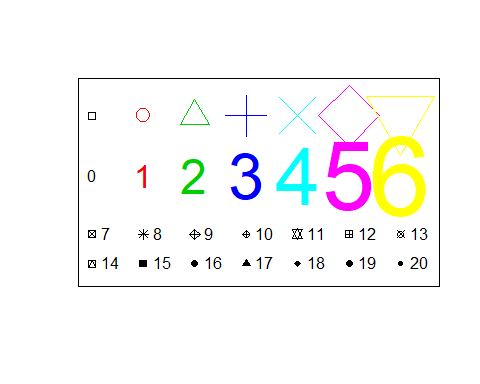
## Q1. The parameters control the size, color and shape of plotting symbol for plot funcions.

## To understand the parameters like cex, col, pch, pos, etc., run the following code one by one.

## Describe the function of these plot arguments. (4 pts.)

## e.g., how the symbol/color/size change

plot(1, 1, xlim=c(1, 7.5), ylim=c(1.75,5), type="n", axes=F, xlab="",  
 ylab="") # Do not plot points  
box()  
points(1:7, rep(4.5, 7), cex=1:7, col=1:7, pch=0:6)  
text(1:7,rep(3.5, 7), labels=paste(0:6), cex=1:7, col=1:7)  
points(1:7,rep(2.5,7), pch=(0:6)+7) # Plot symbols 7 to 13  
text((1:7), rep(2.5,7), paste((0:6)+7), pos=4) # Label with symbol number  
points(1:7,rep(2,7), pch=(0:6)+14) # Plot symbols 14 to 20  
text((1:7), rep(2,7), paste((0:6)+14), pos=4) # Labels with symbol number



### Your answer here

# plot(1, 1, xlim=c(1, 7.5), ylim=c(1.75,5), type=“n”, axes=F, xlab=“”,

# ylab=“”)

# this creates a new plot with an x axis in the range(1,7.5) and y axis in the

# range(1.75,5). type is set to “n” meaning the plot will not produce any points or lines

# axes is set to F meaning the x and y axis calibrations will not be visible

# in the plot. xlab and ylab values are not set so both axes will not have labels

# box()

# This function draws a box around the plot

# points(1:7, rep(4.5, 7), cex=1:7, col=1:7, pch=0:6)

# this function creates 7 points on the plot.The col parameter determines the number(range)of colors used

# in creating the points and the pch parameter determines the symbols that will be used in creating the points

# text(1:7,rep(3.5, 7), labels=paste(0:6), cex=1:7, col=1:7)

# this creates 7 figures(text symbols) labelled 0,1,2,3,4,5,6 pasted in 7 different colors

# points(1:7,rep(2.5,7), pch=(0:6)+7)

# this creates 7 new symbols in the plot ,symbols 7 to 13 to be precise

# text((1:7), rep(2.5,7), paste((0:6)+7), pos=4)

# This pasted 7 new figures: text 7,8,9,10,11,12,13 on the plot positioned beside the plot symbols 7-13.

# points(1:7,rep(2,7), pch=(0:6)+14)

# this created 7 new symbols , plot symbols 14- 20

# text((1:7), rep(2,7), paste((0:6)+14), pos=4)

# this pasted 7 text figures 14,15,16,17,18,19,20 on the plot positioned beside plot symbols 14-20

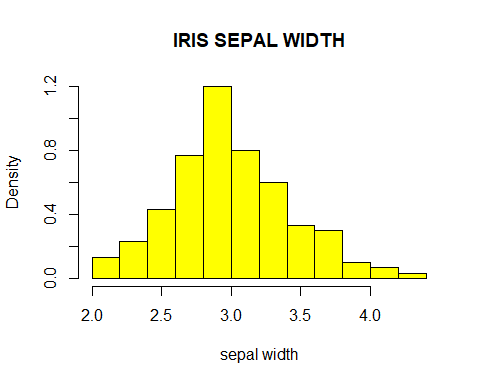
# Part II: Plotting Functions (16 pts.)

## We will discuss plots with built-in iris dataset which you have already seen before.

## Q1. Make a histogram for the variable “Sepal.Width” to show density instead of frequency.(2 pt.)

### Your code here

data(iris)  
hist(iris$Sepal.Width, col = "yellow",prob = TRUE, xlab = "sepal width", main = "IRIS SEPAL WIDTH")

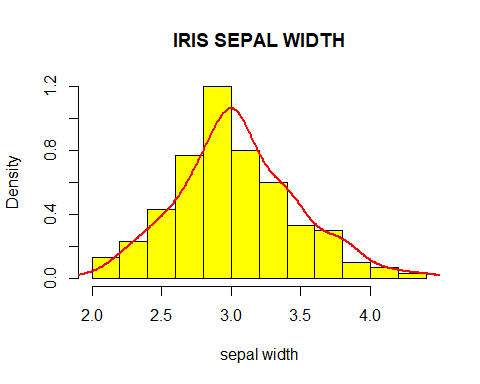


## Q2. Add density curve on the histogram from the previous question.(2 pts.)

## Hint: Adjust arguments of line() to make the line stand out.

### Your code here

hist(iris$Sepal.Width, col="yellow", prob = TRUE, xlab = "sepal width",main = "IRIS SEPAL WIDTH")  
lines(density(iris$Sepal.Width),lwd =2,col = "red")



## Q3. Add a normal distribution curve on the plot from the previous question

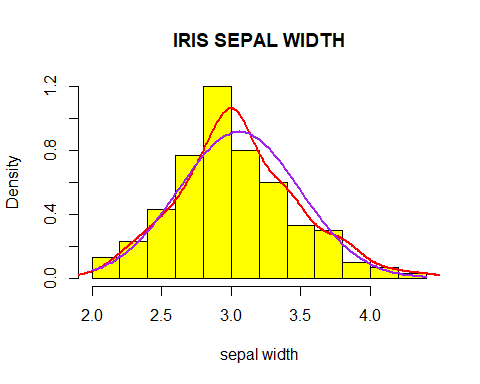
## with mean and standard deviation of Sepal.Width. (2 pts.)

## Compare two curves and discuss the distribution of Sepal.Width. (2 pts.)

## Hint: curve() function may help.

### Your code here

mean\_s.width= mean(iris$Sepal.Width)  
sd\_s.width = sd(iris$Sepal.Width)  
hist(iris$Sepal.Width, col="yellow", prob = TRUE, xlab = "sepal width",main = "IRIS SEPAL WIDTH")  
lines(density(iris$Sepal.Width),lwd =2,col = "red")  
curve(dnorm(x, mean=mean(iris$Sepal.Width), sd=sd(iris$Sepal.Width)), add=TRUE,lwd=2, col = "purple")



### Your answer here

# The Sepal.Width distribution seems to follow the usual pattern of a normal distrubution. However, i observe

# that it has a longer upper tail or is skewed right slightly which represent the few flowers that have

# larger sepal widths.

## Q5. Make a horizontal boxplot to show Sepal.Width in different species.

## Add a rug to your plot. (2 pts.)

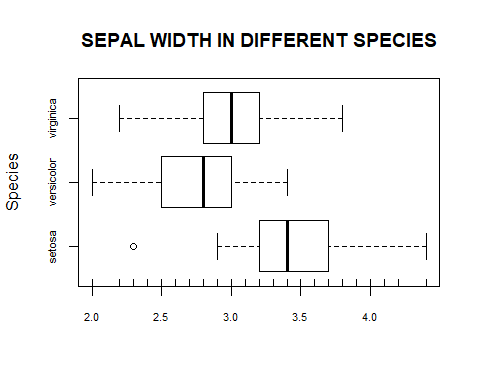
## (1) Try to identify the minimum, lower quartile, median, upper quartile and

## maximum value of virginica’s Sepal.Width by your eyes. (2 pts.)

## (2) Explain what the overlapping in x axis means (2 pts.)

### Your code here

boxplot(Sepal.Width~Species,data=iris, main="SEPAL WIDTH IN DIFFERENT SPECIES", Xlab="Sepal.Width",ylab="Species",horizontal =TRUE, par(cex.axis=0.7))  
rug(iris$Sepal.Width)



### Your answer here

# 1)minimum value = 2.2

# lower quartile = 2.8

# median = 3.0

# upper quartile = 3.2

# maximum value = 3.8

# 2) There could be overlapping in the x axis in terms of the axis labels and tick labels but this can be solved by modifying the

# graphical parameters, however that is not the case with this plot as there is no overlap in the x axis.

## Q6. Make a scatter plot, Petal.Length vs. Petal.Width.

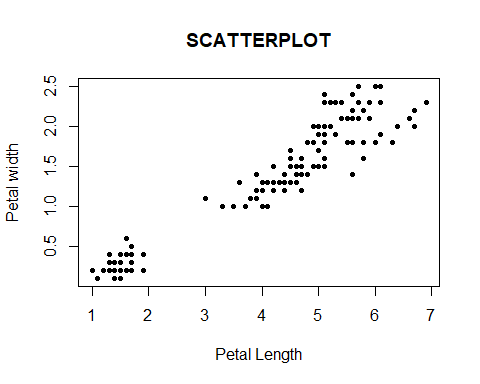
## Mark in the points by species. (2 pts)

## Hint: You could adjust pch argument.

## For example: pch=21 for filled circles, pch=22 for filled squares, pch=23 for filled diamonds.

### Your code here

plot(iris$Petal.Length,iris$Petal.Width,main="SCATTERPLOT", xlab="Petal Length", ylab="Petal width", pch=20)



# Part III. ggplot (10 + 4 pts)

## This question implements ggplot.

## ggplot2 is very powerful, because you can create new graphics

## that are precisely tailored to your problem.

## Google can be helpful if you find yourself stuck, as well as office hours.

### First of all, install ggplot2 and access it.

#install.packages("ggplot2")  
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.4.4

## We will play with plots using diamonds dataset.

## Q1. Use ggplot to make a histogram for the price variable and fill the histogram by the cut Variable.

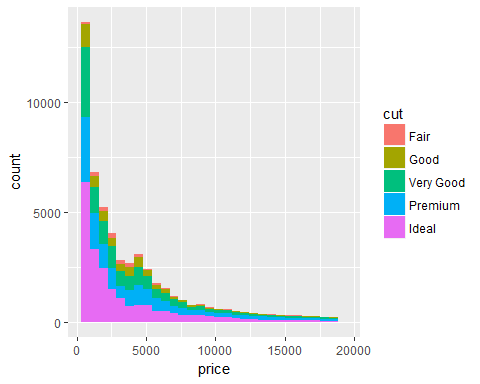
## The plot is similar to

## <http://ggplot2.tidyverse.org/reference/geom_freqpoly-9.png> (2 pts)

### Your code here

data(diamonds)  
ggplot(diamonds, aes(x = price, fill =cut)) + geom\_histogram()

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

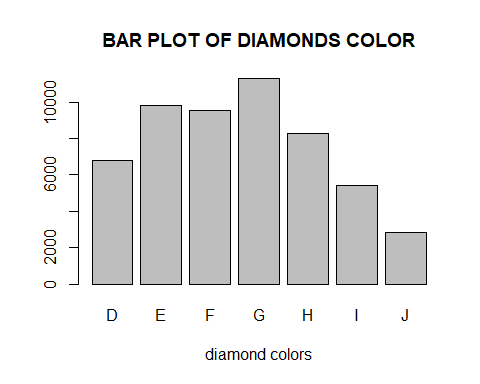


## Q2. Make a bar plot for color of diamonds

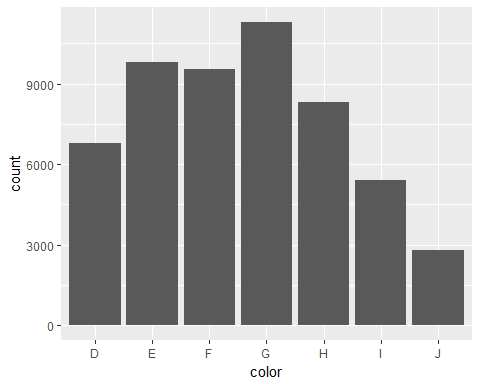
## with both basic plotting function (requried) and ggplot2 (bonus) (2+2 pts.).

### Your code here

diamonds\_color = table(diamonds$color)  
barplot(diamonds\_color, main="BAR PLOT OF DIAMONDS COLOR", xlab="diamond colors")



ggplot(diamonds, aes(x = color)) + geom\_bar()



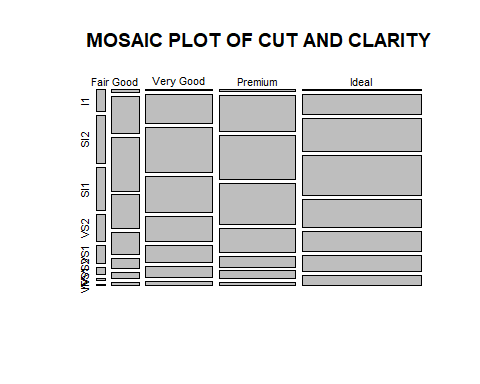
## Q3. Make a mosaic plot with cut and clarity variables.

## You can use a basic plotting function (required) and ggplot2 (bonus).

## To make a mosaic plot with ggplot, you will need the ggmosaic package. (2+2 pts.)

### Your code here

mosaicplot(table(diamonds$cut,diamonds$clarity), main = "MOSAIC PLOT OF CUT AND CLARITY")



#install.packages("ggmosaic")  
library(ggmosaic)

## Warning: package 'ggmosaic' was built under R version 3.4.4

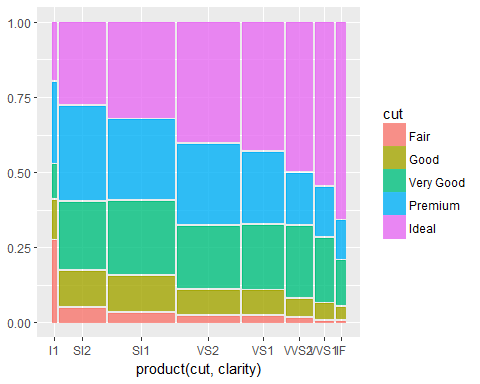
## Loading required package: productplots

## Warning: package 'productplots' was built under R version 3.4.4

##   
## Attaching package: 'ggmosaic'

## The following objects are masked from 'package:productplots':  
##   
## ddecker, hspine, mosaic, prodcalc, spine, vspine

ggplot(data=diamonds) + geom\_mosaic(aes(x=product(cut,clarity), fill=cut))



## Q4. Why is the Data Life Cycle is crucial to understanding

## the opportunities and challenges of making the most of digital data?

## Give two examples. (4 pts.)

# The data life cycle provides a detailed, systematic approach for dealing with data. Every stage of the data life cycle

# corresponds to certain tasks that need to be carried out. While there are many variations of the data life cycle that

# which can be employed, certain important tasks such as data cleaning, analysis are compulsory to ensure that data is correctly utilized to generate insights

# Also the data life cycle needs to be employed in a thorough manner because this ensures that documentation is available for future use , further exploration and discovery.