1. Using the iris dataset...

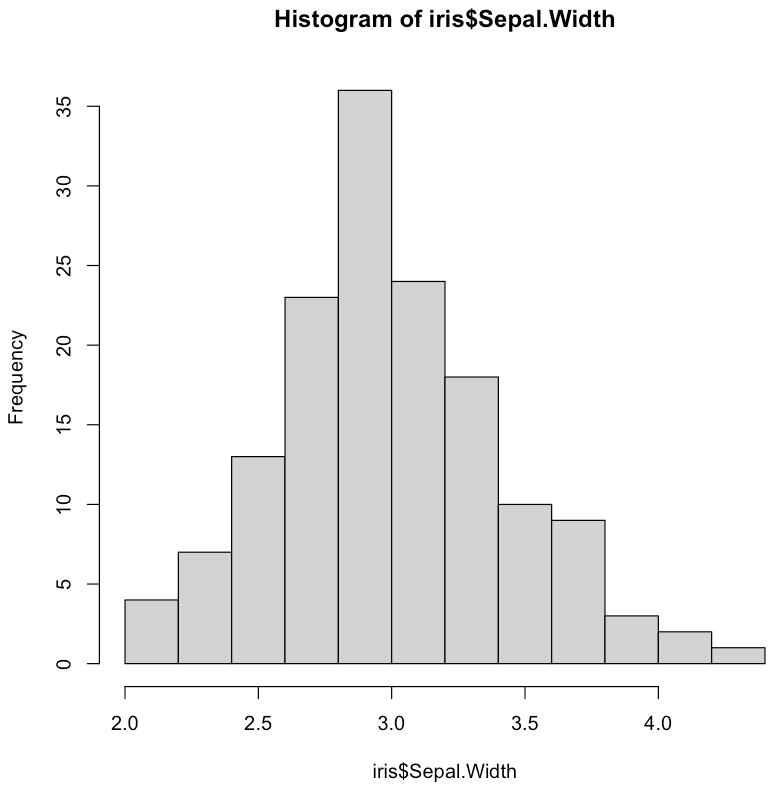
In [5]:
```
head(iris,n = 5)
```

A data.frame: 5 x 5

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
	<dbl>	<dbl>	<dbl>	<dbl>	<fct>
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa

a) Make a histogram of the variable Sepal.Width.

In [7]:
```
par(bg = "white")
hist(iris$Sepal.Width)
```



b) Based on the histogram from #1a, which would you expect to be higher, the mean or the median? Why?

While the graph at a quick glance seems symmetrical, it appears we have a tail on the right side of the histogram. For that reason I would expect see a right skewed graph where the mean is greater than the median, but not by much

c) Confirm your answer to #1b by actually finding these values.

In [11]:
```
sepal_width <- iris$Sepal.Width
cat("Mean is: ", mean(sepal_width), "\n")
cat("Median is: ", median(sepal_width))
```

```
Mean is:  3.057333
Median is:  3
```

d) Only 27% of the flowers have a Sepal.Width higher than _____ cm.

In [27]:
```
# Because we want the last 27% of the data we will want to do 1 - .27 to get
value <- 1 - .27
x <- quantile(iris$Sepal.Width, value)
cat("Only 27% of the flowers have a Sepal.Width higher than " , x, "cm")

# Quick check against data... does answer feel right?
iris$Sepal.Width
```

```
Only 27% of the flowers have a Sepal.Width higher than  3.3 cm
```

3.5 · 3 · 3.2 · 3.1 · 3.6 · 3.9 · 3.4 · 3.4 · 2.9 · 3.1 · 3.7 · 3.4 · 3 · 3 · 4 · 4.4 · 3.9 · 3.5 · 3.8 · 3.8 · 3.4 · 3.7 · 3.6 · 3.3 · 3.4 · 3 · 3.4 · 3.5 · 3.4 · 3.2 · 3.1 · 3.4 · 4.1 · 4.2 · 3.1 · 3.2 · 3.5 · 3.6 · 3 · 3.4 · 3.5 · 2.3 · 3.2 · 3.5 · 3.8 · 3 · 3.8 · 3.2 · 3.7 · 3.3 · 3.2 · 3.2 · 3.1 · 2.3 · 2.8 · 2.8 · 3.3 · 2.4 · 2.9 · 2.7 · 2 · 3 · 2.2 · 2.9 · 2.9 · 3.1 · 3 · 2.7 · 2.2 · 2.5 · 3.2 · 2.8 · 2.5 · 2.8 · 2.9 · 3 · 2.8 · 3 · 2.9 · 2.6 · 2.4 · 2.4 · 2.7 · 2.7 · 3 · 3.4 · 3.1 · 2.3 · 3 · 2.5 · 2.6 · 3 · 2.6 · 2.3 · 2.7 · 3 · 2.9 · 2.9 · 2.5 · 2.8 · 3.3 · 2.7 · 3 · 2.9 · 3 · 3 · 2.5 · 2.9 · 2.5 · 3.6 · 3.2 · 2.7 · 3 · 2.5 · 2.8 · 3.2 · 3 · 3.8 · 2.6 · 2.2 · 3.2 · 2.8 · 2.8 · 2.7 · 3.3 · 3.2 · 2.8 · 3 · 2.8 · 3 · 2.8 · 3.8 · 2.8 · 2.8 · 2.6 · 3 · 3.4 · 3.1 · 3 · 3.1 · 3.1 · 3.1 · 2.7 · 3.2 · 3.3 · 3 · 2.5 · 3 · 3.4 · 3

e) Make scatterplots of each pair of the numerical variables in iris (There should be 6 pairs/plots).

In [33]:
```
# Check Numerical Variables
str(iris)
```

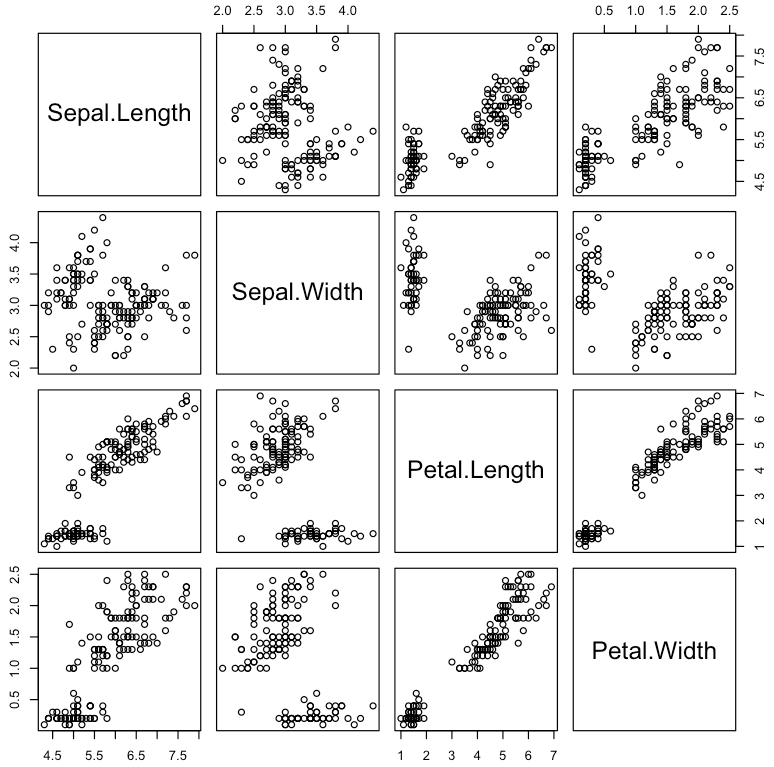
```
'data.frame':   150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1
1 1 1 ...
```

In [39]:
```
par(bg="white")
pairs(iris[, c("Sepal.Length", "Sepal.Width", "Petal.Length", "Petal.Width")])
```

```
# To answer F
cor(iris[, c("Sepal.Length", "Sepal.Width", "Petal.Length", "Petal.Width")])
```

A matrix: 4 x 4 of type dbl

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	1.0000000	-0.1175698	0.8717538	0.8179411
Sepal.Width	-0.1175698	1.0000000	-0.4284401	-0.3661259
Petal.Length	0.8717538	-0.4284401	1.0000000	0.9628654
Petal.Width	0.8179411	-0.3661259	0.9628654	1.0000000



f) Based on #1e, which two variables appear to have the strongest relationship? And which two appear to have the weakest relationship?

Strongest: Petal.Length & Petal.Width
Weakest: Sepal.Length & Sepal.Width

2. Using the PlantGrowth dataset...

In [79]:
```
head(PlantGrowth, n = 5)
str(PlantGrowth)
```

A data.frame: 5 x 2

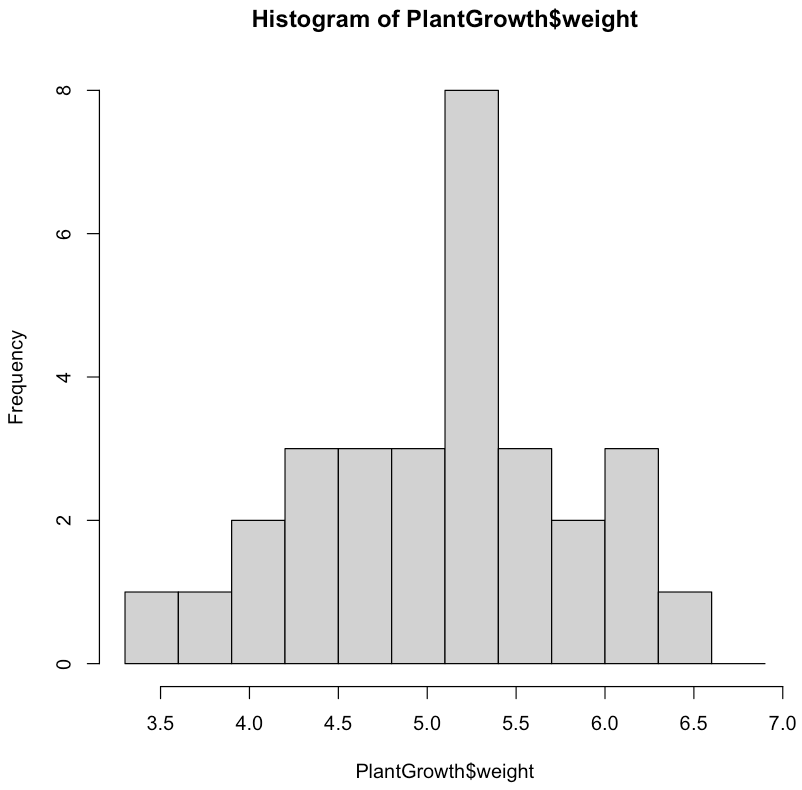
	weight	group
	<dbl>	<fct>
1	4.17	ctrl
2	5.58	ctrl
3	5.18	ctrl
4	6.11	ctrl
5	4.50	ctrl

```
'data.frame':   30 obs. of  2 variables:
 $ weight: num  4.17 5.58 5.18 6.11 4.5 4.61 5.17 4.53 5.33 5.14 ...
 $ group : Factor w/ 3 levels "ctrl","trt1",..: 1 1 1 1 1 1 1 1 1 1 ...
```

a) Make a histogram of the variable weight with breakpoints (bin edges) at every 0.3 units, starting at 3.3.

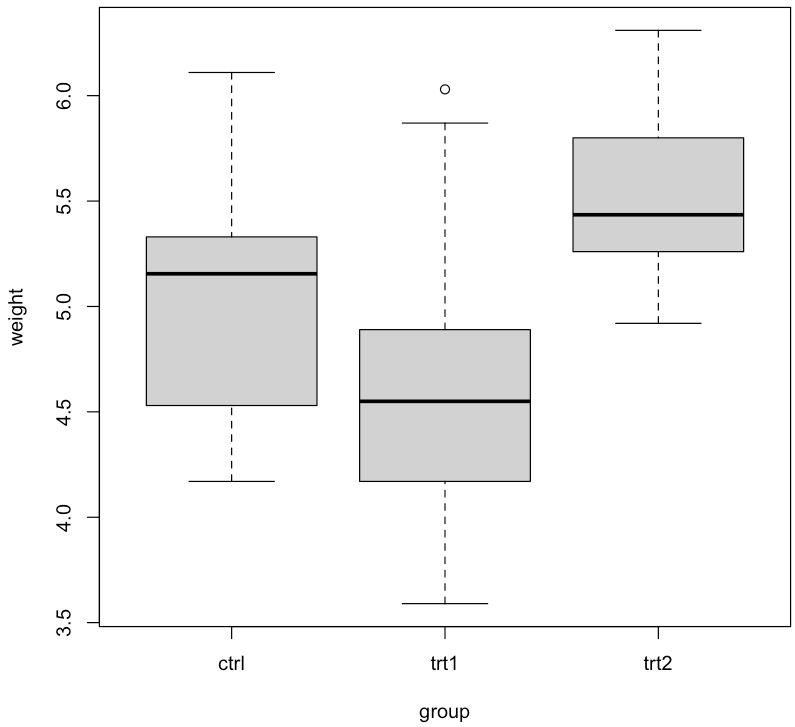
In [59]:
```
par(bg = "white")

hist(PlantGrowth$weight,
     breaks = seq(3.3, ceiling(max(PlantGrowth$weight)), by = 0.3)
)
```



b) Make boxplots of weight separated by group in a single graph.

In [72]:
```
par(bg = "white")
boxplot(weight ~ group, data = PlantGrowth)
```



c) Based on the boxplots in #2b, approximately what percentage of the "trt1" weights are below the minimum "trt2" weight?

Answer: 75%. The minimum for trt2 is the bottom whisker. The bottom of the whisker for trt2 is almost aligned with the top of the box on trt1. The top of a box on a box and whisker chart would be the 3rd Quartile or 75% of the data. We can say this because all of the box and whiskers share the same Y-Axis.

d) Find the exact percentage of the "trt1" weights that are below the minimum "trt2" weight.

Answer: 80%

In [114…
```
# Get min of trt2 data
trt2_min <- min(PlantGrowth$weight[PlantGrowth$group == "trt2"])
cat("Minimum trt2 value", trt2_min , "\n")

# Get Vector of trt1 data
trt1_data <- PlantGrowth$weight[PlantGrowth$group == "trt1"]
cat("All trt1 values are:", trt1_data, "\n")

# Get length of trt1 data
total_length_trt1 <- length(trt1_data)
cat("Total length of all trt1 values is", total_length_trt1, "\n")
```

```r
# Filter trt1 data where the weights are below the min of trt2 data
trt1_values_below_min_trt2 <- trt1_data[trt1_data < trt2_min]
cat("trt1 values that are less than the min of trt2 values are:", trt1_value

# Get the length of that vector
length_filtered_trt1_values <- length(trt1_values_below_min_trt2)
cat("Number of values filtered", length_filtered_trt1_values, "\n")

# Get Percentage
answer <- length_filtered_trt1_values / total_length_trt1 * 100
cat("Percentage is", answer, "%")
```

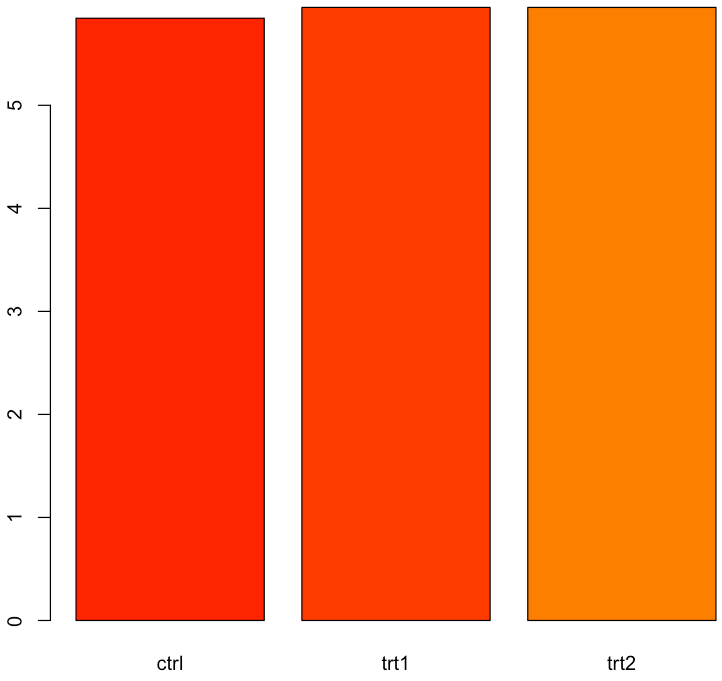
```
Minimum trt2 value 4.92
All trt1 values are: 4.81 4.17 4.41 3.59 5.87 3.83 6.03 4.89 4.32 4.69
Total length of all trt1 values is 10
trt1 values that are less than the min of trt2 values are: 4.81 4.17 4.41 3.
59 3.83 4.89 4.32 4.69
Number of values filtered 8
Percentage is 80 %
```

e) Only including plants with a weight above 5.5, make a barplot of the variable group. Make the barplot colorful using some color palette (in R, try running ?heat.colors and/or check out https://www.r-bloggers.com/palettes-in-r/).

In [140…

```r
filtered_df <- PlantGrowth[PlantGrowth$weight > 5.5, ]
aggr_df <- aggregate(weight ~ group, data = filtered_df, FUN = mean)

par(bg = "white")
barplot(aggr_df$weight,
        names.arg = aggr_df$group,
        col=heatcols)
```


0
1
2
3
4
5
ctrl
trt1
trt2