Data Visualization in R

L. Torgo

ltorgo@fc.up.pt

Faculdade de Ciências / LIAAD-INESC TEC, LA
Universidade do Porto

Oct, 2014

[dcc]
DEPARTAMENTO DE CIÊNCIA DE COMPUTADORES
FACULDADE DE CIÊNCIAS DA UNIVERSIDADE DO PORTO

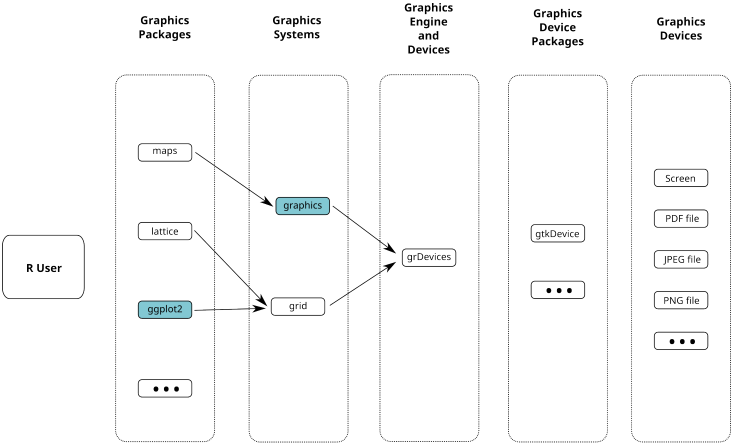
Motivation for Data Visualization

- Humans are outstanding at detecting patterns and structures with their eyes
- Data visualization methods try to explore these capabilities
- In spite of all advantages visualization methods also have several problems, particularly with very large data sets

Outline of what we will learn

- Tools for univariate data
- Tools for bivariate data
- Tools for multivariate data
    - Multidimensional scaling methods

R Graphics


Graphics Packages
Graphics Systems
Graphics Engine and Devices
Graphics Device Packages
Graphics Devices
R User
maps
lattice
ggplot2
• • •
graphics
grid
grDevices
gtkDevice
• • •
Screen
PDF file
JPEG file
PNG file
• • •

Standard Graphics (the `graphics` package)

- R standard graphics, available through package **graphics**, includes several functions that provide standard statistical plots, like:
    - Scatterplots
    - Boxplots
    - Piecharts
    - Barplots
    - etc.
- These graphs can be obtained tyipically by a single function call
    - Example of a scatterplot

```
plot(1:10,sin(1:10))
```

- These graphs can be easily augmented by adding several elements to these graphs (lines, text, etc.)

Standard Graphics (the `graphics` package)

- R standard graphics, available through package **graphics**, includes several functions that provide standard statistical plots, like:
  - Scatterplots
  - Boxplots
  - Piecharts
  - Barplots
  - etc.
- These graphs can be obtained tyipically by a single function call
  - Example of a scatterplot

```
plot(1:10,sin(1:10))
```

- These graphs can be easily augmented by adding several elements to these graphs (lines, text, etc.)

Standard Graphics (the `graphics` package)

- R standard graphics, available through package **graphics**, includes several functions that provide standard statistical plots, like:
  - Scatterplots
  - Boxplots
  - Piecharts
  - Barplots
  - etc.
- These graphs can be obtained tyipically by a single function call
  - Example of a scatterplot

```
plot(1:10,sin(1:10))
```

- These graphs can be easily augmented by adding several elements to these graphs (lines, text, etc.)

Graphics Devices

- R graphics functions produce output that depends on the active graphics device
- The default and more frequently used device is the **screen**
- There are many more graphical devices in R, like the **pdf** device, the **jpeg** device, etc.
- The user just needs to open (and in the end close) the graphics output device she/he wants. R takes care of producing the type of output required by the device
- This means that to produce a certain plot on the screen or as a GIF graphics file the R code is exactly the same. You only need to open the target output device before!
- Several devices may be open at the same time, but only one is the active device

Graphics Devices

- R graphics functions produce output that depends on the active graphics device
- The default and more frequently used device is the **screen**
- There are many more graphical devices in R, like the **pdf** device, the **jpeg** device, etc.
- The user just needs to open (and in the end close) the graphics output device she/he wants. R takes care of producing the type of output required by the device
- This means that to produce a certain plot on the screen or as a GIF graphics file the R code is exactly the same. You only need to open the target output device before!
- Several devices may be open at the same time, but only one is the active device

FACULDADE DE CIÊNCIAS UNIVERSIDADE DO PORTO

Graphics Devices

- R graphics functions produce output that depends on the active graphics device
- The default and more frequently used device is the **screen**
- There are many more graphical devices in R, like the **pdf** device, the **jpeg** device, etc.
- The user just needs to open (and in the end close) the graphics output device she/he wants. R takes care of producing the type of output required by the device
- This means that to produce a certain plot on the screen or as a GIF graphics file the R code is exactly the same. You only need to open the target output device before!
- Several devices may be open at the same time, but only one is the active device

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Graphics Devices

- R graphics functions produce output that depends on the active graphics device
- The default and more frequently used device is the **screen**
- There are many more graphical devices in R, like the **pdf** device, the **jpeg** device, etc.
- The user just needs to open (and in the end close) the graphics output device she/he wants. R takes care of producing the type of output required by the device
- This means that to produce a certain plot on the screen or as a GIF graphics file the R code is exactly the same. You only need to open the target output device before!
- Several devices may be open at the same time, but only one is the active device

Graphics Devices

- R graphics functions produce output that depends on the active graphics device
- The default and more frequently used device is the **screen**
- There are many more graphical devices in R, like the **pdf** device, the **jpeg** device, etc.
- The user just needs to open (and in the end close) the graphics output device she/he wants. R takes care of producing the type of output required by the device
- This means that to produce a certain plot on the screen or as a GIF graphics file the R code is exactly the same. You only need to open the target output device before!
- Several devices may be open at the same time, but only one is the active device

Graphics Devices

- R graphics functions produce output that depends on the active graphics device
- The default and more frequently used device is the **screen**
- There are many more graphical devices in R, like the **pdf** device, the **jpeg** device, etc.
- The user just needs to open (and in the end close) the graphics output device she/he wants. R takes care of producing the type of output required by the device
- This means that to produce a certain plot on the screen or as a GIF graphics file the R code is exactly the same. You only need to open the target output device before!
- Several devices may be open at the same time, but only one is the active device

A few examples

A scatterplot

```
plot(seq(0,4*pi,0.1),sin(seq(0,4*pi,0.1)))
```

The same but stored on a jpeg file

```
jpeg('exp.jpg')
plot(seq(0,4*pi,0.1),sin(seq(0,4*pi,0.1)))
dev.off()
```

And now as a pdf file

```
pdf('exp.pdf',width=6,height=6)
plot(seq(0,4*pi,0.1),sin(seq(0,4*pi,0.1)))
dev.off()
```

Package ggplot2

- Package ggplot2 implements the ideas created by Wilkinson (2005) on a grammar of graphics
- This grammar is the result of a theoretical study on what is a statistical graphic
- ggplot2 builds upon this theory by implementing the concept of a layered grammar of graphics (Wickham, 2009)
- The grammar defines a statistical graphic as:
    - a mapping from data into **aesthetic attributes** (color, shape, size, etc.) of **geometric objects** (points, lines, bars, etc.)

L. Wilkinson (2005). The Grammar of Graphics. Springer.

H. Wickham (2009). A layered grammar of graphics. Journal of Computational and Graphical Statistics.

The Basics of the Grammar of Graphics

- Key elements of a statistical graphic:
  - data
  - aesthetic mappings
  - geometric objects
  - statistical transformations
  - scales
  - coordinate system
  - faceting

Aesthetic Mappings

- Controls the relation between data variables and graphic variables
  - map the *Temperature* variable of a data set into the *x* variable in a scatter plot
  - map the *Species* of a plant into the *colour* of dots in a graphic
  - etc.

Geometric Objects

- Controls what is shown in the graphics
  - show each observation by a point using the aesthetic mappings that map two variables in the data set into the *x*, *y* variables of the plot
  - etc.

Statistical Transformations

- Allows us to calculate and do statistical analysis over the data in the plot
  - Use the data and approximate it by a regression line on the *x*, *y* coordinates
  - Count occurrences of certain values
  - etc.

Scales

- Maps the data values into values in the coordinate system of the graphics device

Coordinate System

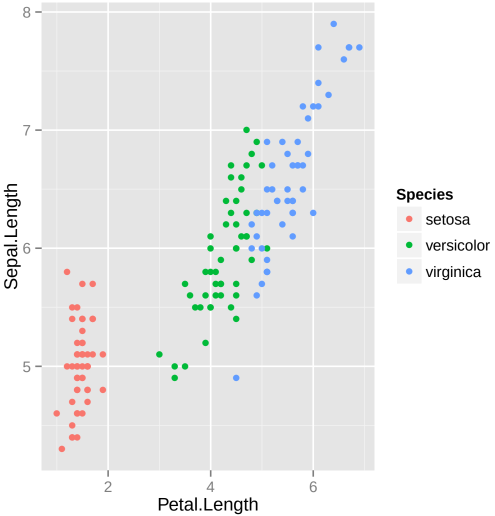
- The coordinate system used to plot the data
    - Cartesian
    - Polar
    - etc.

Faceting

- Split the data into sub-groups and draw sub-graphs for each group

A Simple Example

```
data(iris)
library(ggplot2)
ggplot(iris,
       aes(x=Petal.Length,
           y=Sepal.Length,
           colour=Species
           )
       ) + geom_point()
```



The Distribution of Values of Nominal Variables

The Barplot

- The Barplot is a graph whose main purpose is to display a set of values as heights of bars
- It can be used to display the frequency of occurrence of different values of a nominal variable as follows:
  - First obtain the number of occurrences of each value
  - Then use the Barplot to display these counts

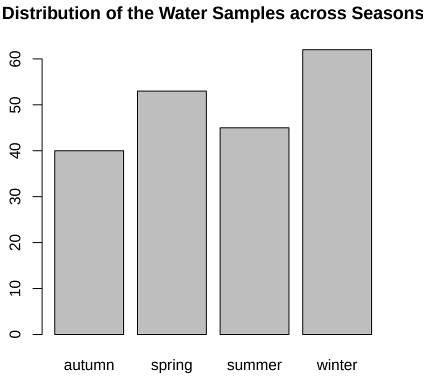
The Distribution of Values of Nominal Variables

The Barplot

- The Barplot is a graph whose main purpose is to display a set of values as heights of bars
- It can be used to display the frequency of occurrence of different values of a nominal variable as follows:
  - First obtain the number of occurrences of each value
  - Then use the Barplot to display these counts

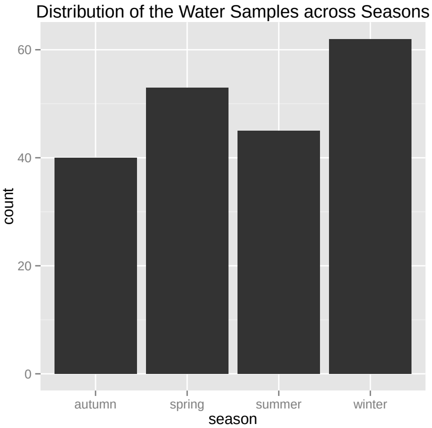
Barplots in base graphics

```
barplot(table(algae$season),
        main='Distribution of the Water Samples across Seasons')
```



Barplots in ggplot2

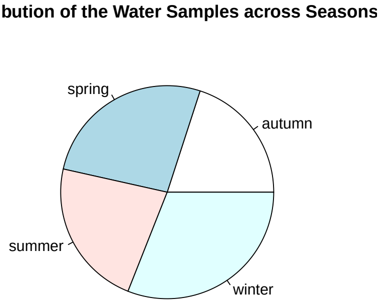
```
ggplot(algae,aes(x=season)) + geom_bar() +
    ggtitle('Distribution of the Water Samples across Seasons')
```



Pie Charts

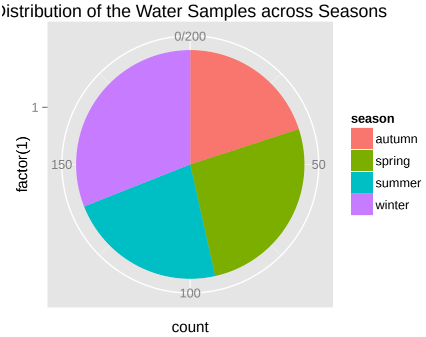
Pie charts serve the same purpose as bar plots but present the information in the form of a pie.

```
pie(table(algae$season),
    main='Distribution of the Water Samples across Seasons')
```



Pie Charts in ggplot

```
ggplot(algae,aes(x=factor(1),fill=season)) + geom_bar(width=1) +
    ggtitle('Distribution of the Water Samples across Seasons') +
        coord_polar(theta="y")
```



The Distribution of Values of a Continuous Variable

The Histogram

- The Histogram is a graph whose main purpose is to display how the values of a continuous variable are distributed
- It is obtained as follows:
    - First the range of the variable is divided into a set of **bins** (intervals of values)
    - Then the number of occurrences of values on each bin is counted
    - Then this number is displayed as a bar

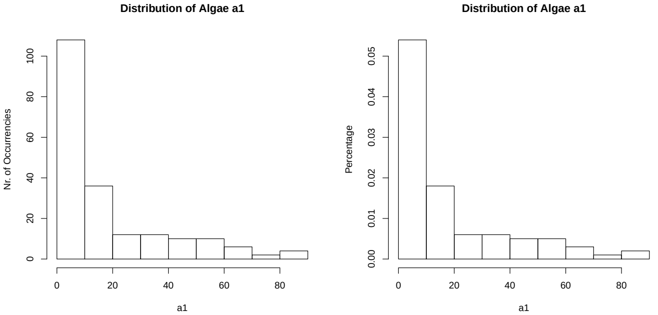
The Distribution of Values of a Continuous Variable

The Histogram

- The Histogram is a graph whose main purpose is to display how the values of a continuous variable are distributed
- It is obtained as follows:
  - First the range of the variable is divided into a set of **bins** (intervals of values)
  - Then the number of occurrences of values on each bin is counted
  - Then this number is displayed as a bar

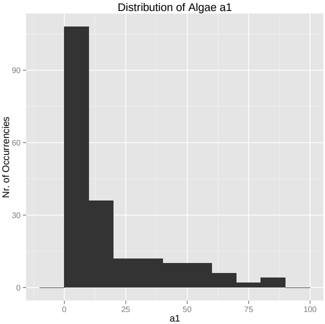
Two examples of Histograms in R

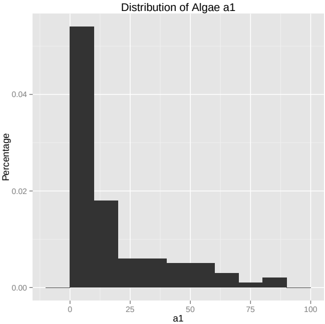
```
hist(algae$a1,main='Distribution of Algae a1',
     xlab='a1',ylab='Nr. of Occurrencies')
hist(algae$a1,main='Distribution of Algae a1',
     xlab='a1',ylab='Percentage',prob=TRUE)
```



Two examples of Histograms in ggplot2

```
ggplot(algae,aes(x=a1)) + geom_histogram(binwidth=10) +
    ggtitle("Distribution of Algae a1") + ylab("Nr. of Occurrencies")
ggplot(algae,aes(x=a1)) + geom_histogram(binwidth=10,aes(y=..density..)) +
    ggtitle("Distribution of Algae a1") + ylab("Percentage")
```


Distribution of Algae a1


Distribution of Algae a1

Problems with Histograms

- Histograms may be misleading in small data sets
- Another key issued is how the limits of the bins are chosen
  - There are several algorithms for this
  - Check the "Details" section of the help page of function `hist()` if you want to know more about this and to obtain references on alternatives
  - Within `ggplot2` you may control this through the `binwidth` parameter

Showing the Distribution of Values

Kernel Density Estimates

- Some of the problems of histograms can be tackled by smoothing the estimates of the distribution of the values. That is the purpose of kernel density estimates
- Kernel estimates calculate the estimate of the distribution at a certain point by smoothly averaging over the neighboring points
- Namely, the density is estimated by

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

where $K()$ is a kernel function and h a bandwidth parameter

Showing the Distribution of Values

Kernel Density Estimates

- Some of the problems of histograms can be tackled by smoothing the estimates of the distribution of the values. That is the purpose of kernel density estimates
- Kernel estimates calculate the estimate of the distribution at a certain point by smoothly averaging over the neighboring points
- Namely, the density is estimated by

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

where $K()$ is a kernel function and h a bandwidth parameter

Showing the Distribution of Values

Kernel Density Estimates

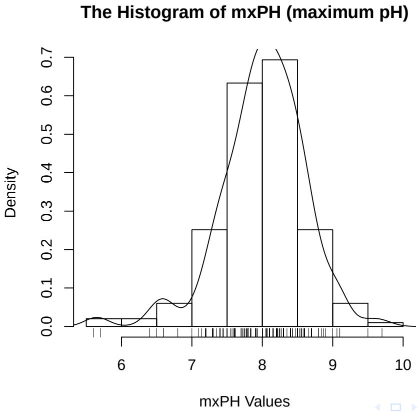
- Some of the problems of histograms can be tackled by smoothing the estimates of the distribution of the values. That is the purpose of kernel density estimates
- Kernel estimates calculate the estimate of the distribution at a certain point by smoothly averaging over the neighboring points
- Namely, the density is estimated by

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

where $K()$ is a kernel function and h a bandwidth parameter

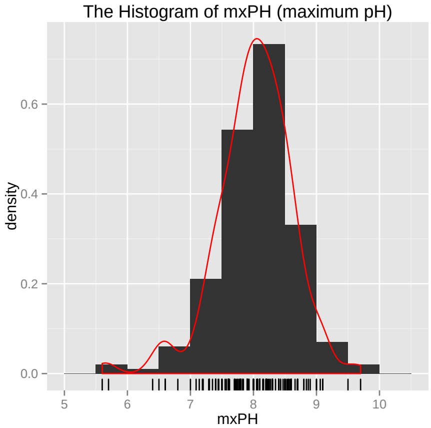
An Example and how to obtain it in R basic graphics

```
hist(algae$mxPH,main='The Histogram of mxPH (maximum pH)',
     xlab='mxPH Values',prob=T)
lines(density(algae$mxPH,na.rm=T))
rug(algae$mxPH)
```



An Example and how to obtain it in ggplot2

```
ggplot(algae,aes(x=mxPH)) + geom_histogram(binwidth=.5, aes(y=..density..)) +
    geom_density(color="red") + geom_rug() + ggtitle("The Histogram of mxPH (maximum pH)")
```



Showing the Distribution of Values

Graphing Quantiles

- The *x* quantile of a continuous variable is the value below which an *x*% proportion of the data is
- Examples of this concept are the 1st (25%) and 3rd (75%) quartiles and the median (50%)
- We can calculate these quantiles at different values of *x* and then plot them to provide an idea of how the values in a sample of data are distributed

Showing the Distribution of Values

Graphing Quantiles

- The *x* quantile of a continuous variable is the value below which an *x*% proportion of the data is
- Examples of this concept are the 1st (25%) and 3rd (75%) quartiles and the median (50%)
- We can calculate these quantiles at different values of *x* and then plot them to provide an idea of how the values in a sample of data are distributed

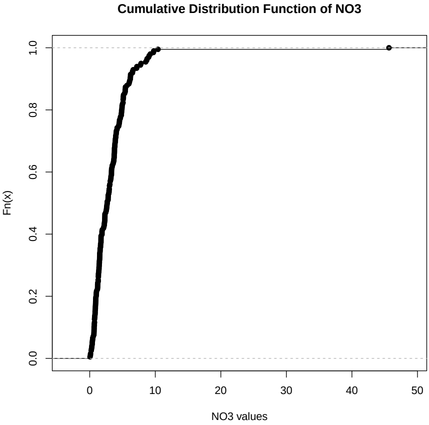
Showing the Distribution of Values

Graphing Quantiles

- The *x* quantile of a continuous variable is the value below which an *x*% proportion of the data is
- Examples of this concept are the 1st (25%) and 3rd (75%) quartiles and the median (50%)
- We can calculate these quantiles at different values of *x* and then plot them to provide an idea of how the values in a sample of data are distributed

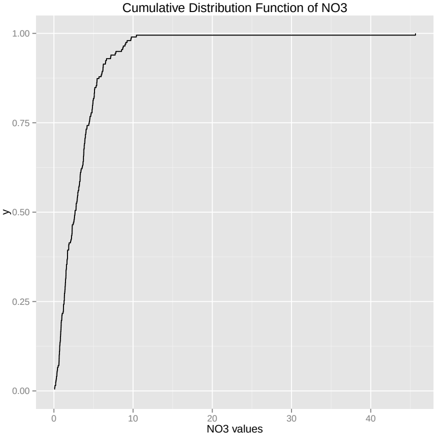
The Cumulative Distribution Function in Base Graphics

```
plot(ecdf(algae$NO3),
     main='Cumulative Distribution Function of NO3',xlab='NO3 values')
```



The Cumulative Distribution Function in `ggplot`

```
ggplot(algae,aes(x=NO3)) + stat_ecdf() + xlab('NO3 values') +
    ggtitle('Cumulative Distribution Function of NO3')
```



QQ Plots

- Graphs that can be used to compare the observed distribution against the Normal distribution
- Can be used to visually check the hypothesis that the variable under study follows a normal distribution
- Obviously, more formal tests also exist

QQ Plots

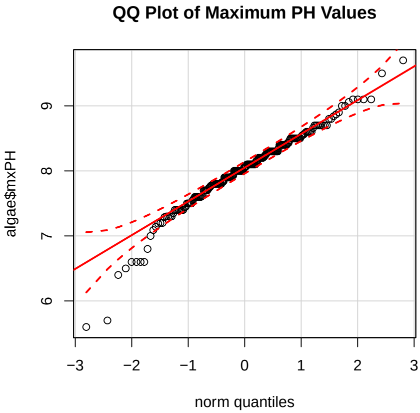
- Graphs that can be used to compare the observed distribution against the Normal distribution
- Can be used to visually check the hypothesis that the variable under study follows a normal distribution
- Obviously, more formal tests also exist

QQ Plots

- Graphs that can be used to compare the observed distribution against the Normal distribution
- Can be used to visually check the hypothesis that the variable under study follows a normal distribution
- Obviously, more formal tests also exist

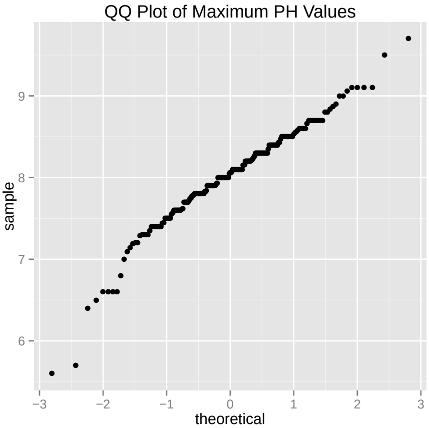
QQ Plots from package `car` using base graphics

```
library(car)
qqPlot(algae$mxPH,main='QQ Plot of Maximum PH Values')
```



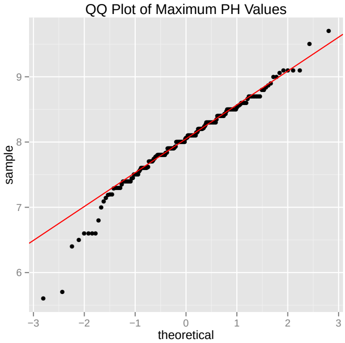
QQ Plots using ggplot2

```
ggplot(algae,aes(sample=mxPH)) + stat_qq() +
    ggtitle('QQ Plot of Maximum PH Values')
```



QQ Plots using `ggplot2` (2)

```
q.x <- quantile(algae$mxPH,c(0.25,0.75),na.rm=TRUE)
q.z <- qnorm(c(0.25,0.75))
b <- (q.x[2] - q.x[1])/(q.z[2] - q.z[1])
a <- q.x[1] - b * q.z[1]
ggplot(algae,aes(sample=mxPH)) + stat_qq() +
    ggtitle('QQ Plot of Maximum PH Values') +
        geom_abline(intercept=a,slope=b,color="red")
```



Showing the Distribution of Values

Box Plots

- Box plots provide interesting summaries of a variable distribution
- For instance, they inform us of the interquartile range and of the outliers (if any)

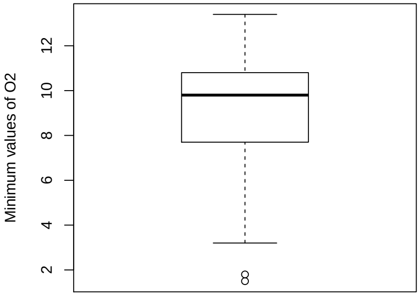
Showing the Distribution of Values

Box Plots

- Box plots provide interesting summaries of a variable distribution
- For instance, they inform us of the interquartile range and of the outliers (if any)

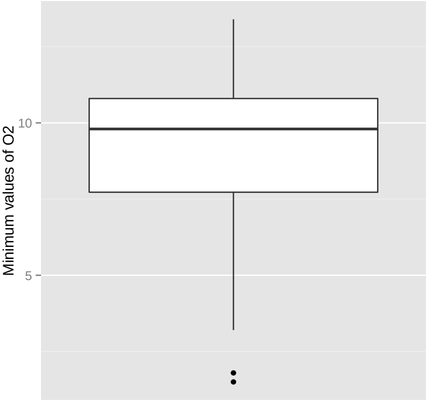
An Example with base graphics

```
boxplot(algae$mnO2,ylab='Minimum values of O2')
```



An Example with base ggplot2

```
ggplot(algae,aes(x=factor(0),y=mnO2)) + geom_boxplot() +
    ylab("Minimum values of O2") + xlab("") + scale_x_discrete(breaks=NULL)
```



Hands on Data Visualization - the Algae data set

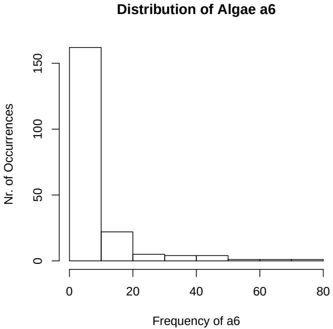
Using the Algae data set from package `DMwR` answer to the following questions:

1. Create a graph that you find adequate to show the distribution of the values of algae `a6` solution
2. Show the distribution of the values of `size` solution
3. Check visually if it is plausible to consider that `oPO4` follows a normal distribution solution

Solution of Exercise 1

- Create a graph that you find adequate to show the distribution of the values of algae `a6`

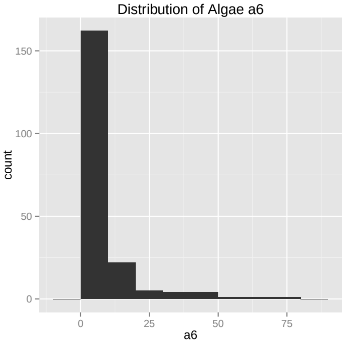
```
hist(algae$a6,main="Distribution of Algae a6",xlab="Frequency of a6",
     ylab="Nr. of Occurrences")
```



Solution of Exercise 1 with ggplot2

- Create a graph that you find adequate to show the distribution of the values of algae a6

```
ggplot(algae,aes(x=a6)) + geom_histogram(binwidth=10) +
    ggtitle("Distribution of Algae a6")
```

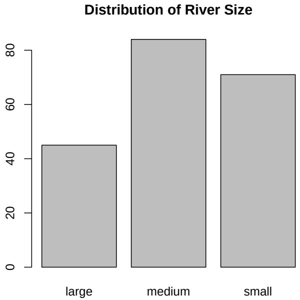


Go Back

Solution to Exercise 2

- Show the distribution of the values of `size`

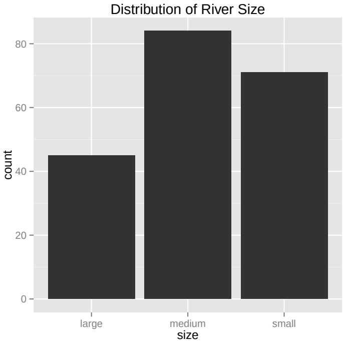
```
barplot(table(algae$size), main = "Distribution of River Size")
```



Solution to Exercise 2 with ggplot2

- Show the distribution of the values of size

```
ggplot(algae,aes(x=size)) + geom_bar() +
    ggtitle("Distribution of River Size")
```

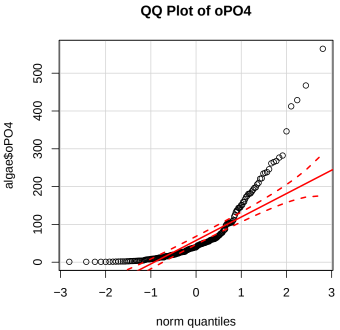


Go Back

Solutions to Exercise 3

- Check visually if it is plausible to consider that `oPO4` follows a normal distribution

```
library(car)
qqPlot(algae$oPO4,main = "QQ Plot of oPO4")
```



Go Back

Conditioned Graphs

- Data sets frequently have nominal variables that can be used to create sub-groups of the data according to these variables values
  - e.g. the sub-group of male clients of a company
- Some of the visual summaries described before can be obtained on each of these sub-groups
- Conditioned plots allow the simultaneous presentation of these sub-group graphs to better allow finding eventual differences between the sub-groups
- Base graphics do not have conditioning but `ggplot2` has it through the concept of facets

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Conditioned Graphs

- Data sets frequently have nominal variables that can be used to create sub-groups of the data according to these variables values
  - e.g. the sub-group of male clients of a company
- Some of the visual summaries described before can be obtained on each of these sub-groups
- Conditioned plots allow the simultaneous presentation of these sub-group graphs to better allow finding eventual differences between the sub-groups
- Base graphics do not have conditioning but `ggplot2` has it through the concept of facets

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Conditioned Graphs

- Data sets frequently have nominal variables that can be used to create sub-groups of the data according to these variables values
  - e.g. the sub-group of male clients of a company
- Some of the visual summaries described before can be obtained on each of these sub-groups
- Conditioned plots allow the simultaneous presentation of these sub-group graphs to better allow finding eventual differences between the sub-groups
- Base graphics do not have conditioning but `ggplot2` has it through the concept of facets

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

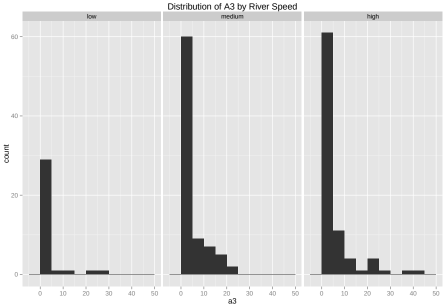
Conditioned Graphs

- Data sets frequently have nominal variables that can be used to create sub-groups of the data according to these variables values
  - e.g. the sub-group of male clients of a company
- Some of the visual summaries described before can be obtained on each of these sub-groups
- Conditioned plots allow the simultaneous presentation of these sub-group graphs to better allow finding eventual differences between the sub-groups
- Base graphics do not have conditioning but `ggplot2` has it through the concept of facets

Conditioned Histograms

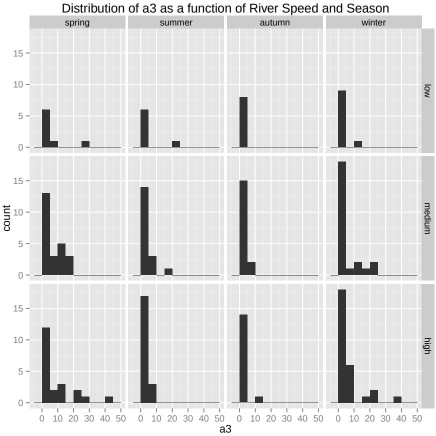
Goal: Constrast the distribution of data sub-groups

```
algae$speed <- factor(algae$speed,levels=c("low","medium","high"))
ggplot(algae,aes(x=a3)) + geom_histogram(binwidth=5) + facet_wrap(~ speed)
    ggtitle("Distribution of A3 by River Speed")
```



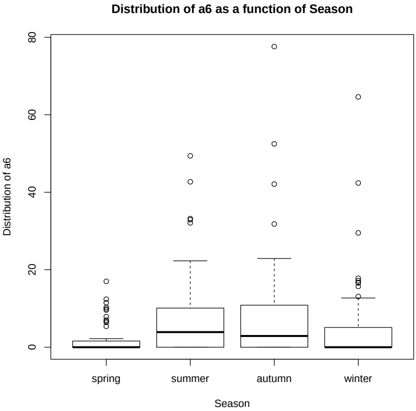
Conditioned Histograms (2)

```
algae$season <- factor(algae$season,levels=c("spring","summer","autumn","winter"))
ggplot(algae,aes(x=a3)) + geom_histogram(binwidth=5) + facet_grid(speed~season) +
    ggtitle('Distribution of a3 as a function of River Speed and Season')
```



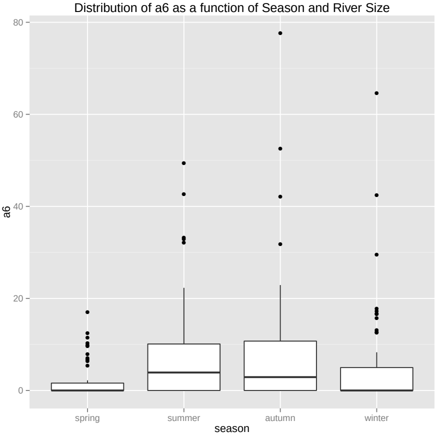
Conditioned Box Plots on base graphics

```
boxplot(a6 ~ season,algae,
    main='Distribution of a6 as a function of Season',
    xlab='Season',ylab='Distribution of a6')
```



Conditioned Box Plots on ggplot2

```
ggplot(algae,aes(x=season,y=a6))+geom_boxplot() +
    ggtitle('Distribution of a6 as a function of Season and River Size')
```



Hands on Data Visualization - Algae data set

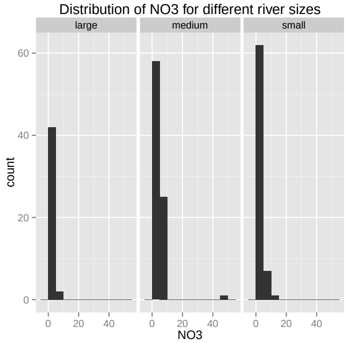
Using the Algae data set from package `DMwR` answer to the following questions:

1. Produce a graph that allows you to understand how the values of NO3 are distributed across the sizes of river solution
2. Try to understand (using a graph) if the distribution of algae a1 varies with the speed of the river solution

Solutions to Exercise 1

- Produce a graph that allows you to understand how the values of NO3 are distributed across the sizes of river.

```
ggplot(algae,aes(x=NO3)) + geom_histogram(binwidth=5) + facet_wrap(~size)
    ggtitle("Distribution of NO3 for different river sizes")
```

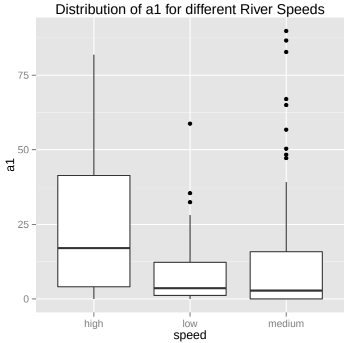


Go Back

Solutions to Exercise 2

- Try to understand (using a graph) if the distribution of algae a1 varies with the speed of the river

```
ggplot(algae,aes(x=speed,y=a1)) + geom_boxplot() +
    ggtitle("Distribution of a1 for different River Speeds")
```

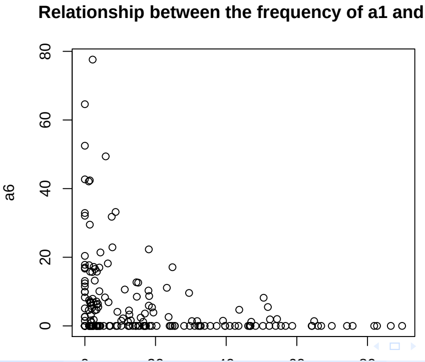


Go Back

Scatterplots in base graphics

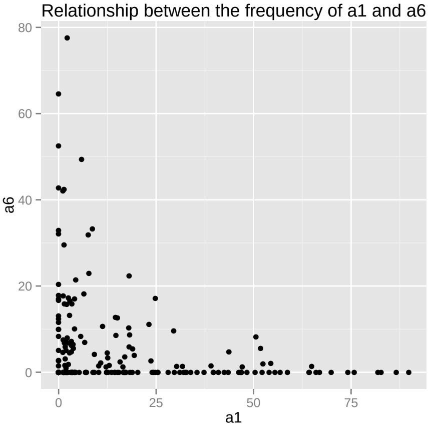
- The Scatterplot is the natural graph for showing the relationship between two numerical variables

```
plot(algae$a1,algae$a6,
     main='Relationship between the frequency of a1 and a6',
     xlab='a1',ylab='a6')
```



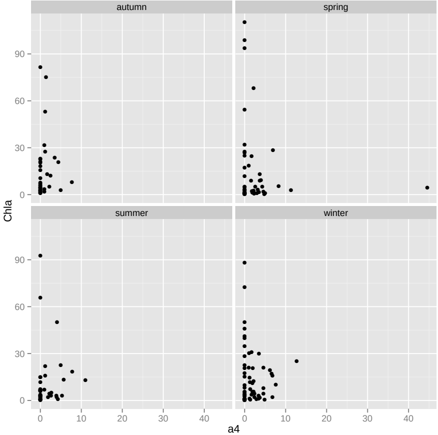
Scatterplots in ggplot2

```
ggplot(algae,aes(x=a1,y=a6)) + geom_point() +
    ggtitle('Relationship between the frequency of a1 and a6')
```



Conditioned Scatterplots using the `ggplot2` package

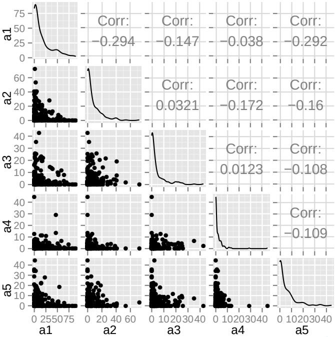
```
ggplot(algae,aes(x=a4,y=Chla)) + geom_point() + facet_wrap(~season)
```



Scatterplot matrices through package `GGally`

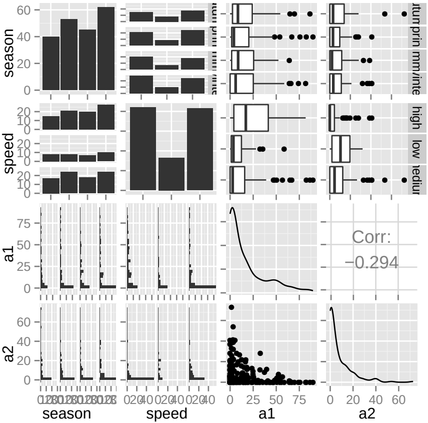
- These graphs try to present all pairwise scatterplots in a data set.
- They are unfeasible for data sets with many variables.

```r
library(GGally)
ggpairs(algae,columns=12:16,
        diag=list(continuous="density", discrete="bar"),axisLabels="show")
```



Scatterplot matrices involving nominal variables

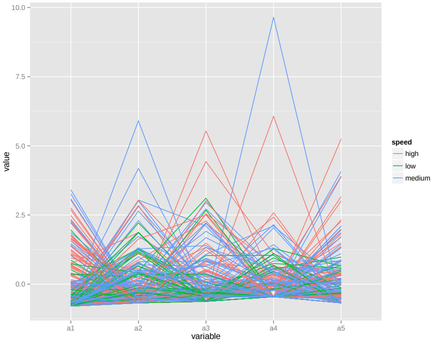
```
ggpairs(algae,columns=c("season","speed","a1","a2"),axisLabels="show")
```



Parallel Plots

- Parallel plots are also interesting for visualizing a data set

```
ggparcoord(algae,columns=12:16,groupColumn="speed")
```


value
a1
a2
a3
a4
a5
variable
speed
high
low
medium